Computers & Geosciences 142 (2020) 104481

Contents lists available at ScienceDirect

Computers and Geosciences

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

Petrographic microfacies classification with deep convolutional neural networks

Rafael Pires de Lima[a,b,*], David Duarte[a], Charles Nicholson[c], Roger Slatt[a], Kurt J. Marfurt[a]

[a] School of Geosciences, The University of Oklahoma, 100 East Boyd Street, RM 710, Norman, OK, 73019, USA
[b] Geological Survey of Brazil – CPRM, 55 Rua Costa, São Paulo, São Paulo, Brazil
[c] School of Industrial and Systems Engineering, The University of Oklahoma, 202 West Boyd Street, RM 124, Norman, OK, 73019, USA

ARTICLE INFO

Keywords:
Petrography thin section analysis
Rock thin section
Convolutional neural networks
Transfer learning

ABSTRACT

Petrographic analysis is based on the microscopic description and classification of rocks and is a crucial technique for sedimentary and diagenetic studies. When compared to hand specimens, thin sections provide better and more accurate means for analysis of mineral proportion, distribution, texture, pore space analysis, and cement composition. Most petrographic analysis relies on visual inspection of rock thin sections under a microscope, a task that is laborious even for experienced geologists. Large projects with a tight time frame requiring the analysis of a large amount of thin sections may require multiple petrographers, thereby risking the introduction of inconsistency in the analysis. To address this challenge, we explore the use of deep convolutional neural networks (CNN) as a tool for acceleration and automatization of microfacies classification. We make use of transfer learning based on robust and reliable CNN models trained with a large amount of non-geological images. With a relatively small number of labeled thin sections used in "fine-tuning" training we are able to adapt CNN models that achieve low error levels (<5%) for the classification of microfacies from the same dataset, and moderate results (<40%) for the classification of microfacies of thin sections from different datasets. These alternate datasets differ from the training data on two independent factors: the thin sections are from different formations and are prepared by different laboratories. While becoming widely accepted as a useful tool in the biological and manufacturing disciplines, CNN is currently underutilized in the geoscience community; we foresee an increase of use of such techniques to help accelerate and quantify a wide variety of geological tasks.

1. Introduction

Petrography focuses on the microscopic description and classification of rocks, and remains one of the most used techniques in geoscience studies. The essential tool in petrographic analysis is an optical microscope that uses plane or polarized transmitted light to capture the optical properties of minerals. Using the optical microscope, the geologist or petrographer examines a rock thin section, which is a flat rock sample usually 30 μm thick, mounted on a glass slide. The goal is to observe and describe the characteristics of the rock such as grain geometry, structure, mineralogical composition, fossil content, and texture. Based on these characteristics, the petrographer defines different rock types called microfacies. Because this study only uses rock samples description at a microscopic level (thin section), we will refer to these rock types as microfacies. Petrographic studies are essential components of geological analysis, ranging from academic studies of mid-ocean ridges to petroleum-industry exploration and development of shale resource plays.

One of the most important uses of petrographic studies is to define microfacies. However, hundreds of thin sections need to be described when classifying microfacies, which is time-consuming. Although the point-count method provides more accurate classification for a thin section, point counts are often discarded as a classification option as it is considered a draining task. In our experience, a qualified geologist can take up to 20 min to count 300 points (the minimum number of points necessary for the point-counting methodology) in a single thin section when the petrographer is familiar with the mineralogical composition of the rock. Due to the long amount of time required for the analysis of a single sample, the mechanical thin section point-count is often replaced by an interpretative approach. A single thin section interpretation can take less than a minute, in cases in which the petrographer is familiar with the microfacies, or up to tens of minutes, in cases in which the thin

* Corresponding author. Geological Survey of Brazil – CPRM, 55 Rua Costa, São Paulo, São Paulo, Brazil.
E-mail address: rafael.lima@cprm.gov.br (R. Pires de Lima).

https://doi.org/10.1016/j.cageo.2020.104481
Received 16 August 2019; Received in revised form 23 March 2020; Accepted 25 March 2020
Available online 10 June 2020

section presents elements that are unfamiliar to the petrographer. The interpretation process can be subjective, thereby enhancing the risk of inconsistent labeling. Cheng et al. (2018) observed that new thin sections are continuously produced, adding to the number of samples that need to be analyzed and archived by the geoscience community. Large amounts of thin sections are constantly analyzed due to the acquisition of new data or reinterpretation of legacy data to ensure consistency.

Launeau and Robin (1996), Přikryl (2001), and Nasseri and Mohanty (2008) reported that the progress of computer-aided image analysis techniques has facilitated the characterization of the microscopic properties of the rock through analysis of digital thin section images. The need to partially automate this process has resulted in the proposal of several machine learning (ML) methods. For example, Sudakov et al. (2019) reported that convolutional neural networks (CNNs) outperformed previous models for permeability prediction using X-ray microtomography. Maitre et al. (2019) used different supervised and unsupervised learning techniques to identify mineral grains in sieved sand samples from natural glacial sediments. Our goal is to generate ML models for the classification of microfacies observed in thin sections that could produce reliable results in a fraction of the time used for manual classification and to provide the possibility for a more quantitative thin section classification analysis.

The microfacies description obtained through images of thin sections are analogous to image classification problems. Datta et al. (2008) reported that image classification is one of the tasks in which machines have excelled, often obtaining faster and more accurate results than humans. Because ML models have been successful in a wide variety of image classification problems, we test CNNs for microfacies classification of thin section photographs.

We begin our paper with a brief review of recent advances in using CNN as image classification in other fields, as well as some of the limited CNN applications using rock thin section data. Next, we describe the thin section preparation and data. We then describe the processing and analysis performed on the data and summarize our results. We conclude with a summary of the advantages and limitations of the technology.

1.1. A short review of image processing using machine learning

Customary ML methods are limited in their ability to process raw data (such as the pixel values of an image). Due to such limitations, for many years the construction of a pattern-recognition model demanded carefully detailed feature engineering (e.g. the analysis of the wings of an insect or the leaves of a tree) performed by domain experts (LeCun et al., 2015; Yin et al., 2017). Yang et al. (2018) observed that one of the reasons deep learning (DL) models attracted the attention of the research community is DL's capacity to discover an effective feature transformation for a specific task. Current progress in DL models, specifically CNN architectures, are the new the state-of-the-art in visual object recognition and detection, speech recognition, and many other fields of study (LeCun et al., 2015). The model described by Krizhevsky et al. (2012), frequently referenced to as AlexNet, is considered a breakthrough and influenced the rapid adoption of DL in the computer vision field (LeCun et al., 2015). A variant of AlexNet won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC, Russakovsky et al., 2015) in 2012 achieving a top-5 test error rate (how often a true label is not one of the top 5 labels assigned by the model, a common metric for the ILSVRC) of 15%. The second-best entry for ILSVRC in 2012 had a top-5 error rate of 26%. AlexNet, with only five convolutional layers, has 60 million parameters to be trained. At first glance, such a large number of parameters might seem like a drawback for the implementation of DL models. However, with the advances of graphics processing units (GPUs), the previously prohibitive long training time has been significantly reduced (Mou et al., 2017; Yang et al., 2018).

In 2012 AlexNet used a five-layer deep CNN model; today many models competing in the ILSVRC use twenty to hundreds of layers. Huang et al. 2016 has even proposed models with thousands of layers. Due to the vast number of operations performed in deep CNN models, it is often difficult to discuss the interpretability, or the degree to which a decision made by a model can be rationalized. For this reason, many workers consider CNN to be a black box, with CNN interpretability itself a research topic (e. g. Simonyan et al., 2013; Olah et al., 2017, 2018; Yin et al., 2017).

Recent CNN developments include several model architectures that achieved top-5 error rates under 10% in the ILSVRC dataset (e.g. Szegedy et al., 2014; Chollet, 2016; He et al., 2016a; Huang, Liu, et al., 2016; Sandler et al., 2018). Yosinski et al. (2014) and Yin et al. (2017) also reported that the parameters learned by the layers in many CNN models trained on images exhibit a very common behavior. The layers closer to the input data tend to learn general features, such as edge detection/enhancement filters or color blobs. Then there is a transition to more specific dataset features, such as faces, feathers, or object parts. These general-specific CNN layers properties that lead to the development of transfer learning (e.g. Caruana, 1995; Bengio, 2012; Yosinski et al., 2014).

In transfer learning, first a CNN model is trained on a dataset for a primary task using large amounts of data. After training, the weights of the model are then repurposed or transferred to a second CNN that can be trained using a smaller dataset, generally domain-specific, for a secondary task (Yosinski et al., 2014).The domain-specific characteristics of a CNN being used for a new task are often addressed through fine-tuning. We provide a brief explanation of the fine-tuning process in the Methods section. Carranza-Rojas et al. (2017) observed that the processes of transfer learning and fine-tuning are important tools that can be used to address the shortage of sufficient domain-specific training data.

Even though large datasets help the performance of DL models, the combination of these technologies (CNNs, transfer learning, and fine-tuning) facilitated the application of DL techniques to other scientific fields. Cunha et al. (2020) applied transfer learning to highlight faults on seismic volumes, Carranza-Rojas et al. (2017) used transfer learning for herbarium specimen classification, Esteva et al. (2017) for dermatologist-level classification of skin cancer, Gomez Villa et al. (2017) for camera-trap images, Hong et al. (2018) for soccer video scene and event classification, Chen et al. (2018) for airplane detection using remote sensing images, and Pires de Lima et al. (2019b) for oil field drill core images. In a study analyzing medical image data, Qayyum et al. (2017) found that transfer learning achieved results comparable to, or better than, results from training a CNN model with randomly initialized parameters. Given this record of success to diverse applications, we hypothesize that ML models will also be beneficial for thin section microfacies classification.

1.2. Machine learning for petrographic image classification

Cheng and Guo (2017) used CNN models to perform image classification based on granularity analysis from thin section images. The authors successfully differentiated between three feldspathic sandstone classes based solely on grain size: coarse-grained, medium-grained, and fine-grained rocks, achieving an accuracy of 98.5%. With high-resolution micro-computed tomography images or rock samples, Karimpouli and Tahmasebi (2019) used CNN to perform the segmentation of minerals in images mainly composed of quartz. Cheng et al. (2018) used CNN for the image retrieval of rock thin sections. The CNN is used to extract features from the thin section images which are then stored in a feature database. The images can then be retrieved based on estimates of the similarity between different images, those thin section images stored in the database and the new thin section image to be classified. Pires de Lima et al. (2019a) presented some preliminary results of geoscientific images classifications, including thin section images. Budennyy et al. (2017) used image processing and ML to perform a semi-automatic calculation of thin section minerals key features of thin section minerals, such as grain rugosity and roundness.

Huang et al. (2016) noted that when crafting CNN models, researchers are uncertain whether to choose from shorter or deeper networks. Shorter networks have a more efficient forward and backward information flow; however, they might not be expressive enough to represent the image features properly. Deeper networks can generate more complex models, helping in feature extraction, but are more difficult to train in practice (e.g., due to computational costs, slower convergence, vanishing and exploding gradients among others). We avoid the challenges of model architecture development making use of well-established and robust CNN models previously trained on the ILSVRC.

1.3. Data

In our study, we analyze 98 thin sections under plane polarized light (PPL) to classify them within five microfacies: argillaceous siltstone, bioturbated siltstones, calcareous siltstone, porous calcareous siltstones, and massive calcite-cemented siltstones. All these microfacies can be identified using plane polarized light and a 10X magnification zoom. We take three randomly-placed photographic images for every thin section. Table 1 summarizes the number of thin sections and respective photographs taken for each one of the five microfacies. The thin sections were acquired from five cores from the Sycamore Formation (Early Mississippian strata) in Carter County and Stephens County, Oklahoma. This is the main data of the study and is separated in training, validation, and test sets.

To further evaluate whether the models generated from the data in Table 1 have more general applicability in classifying thin sections from different geologic formations and processed by different laboratories, we use thin section images from the public domain (referred to as public data) coming from Sycamore and Meramec formations (both Early Mississippian strata) stored at the Oklahoma Petroleum Information Center (OPIC) (Table 2). Early Mississippian strata in the Anadarko Basin in Oklahoma consist of a mixed carbonate-siliciclastic system. Slight differences in lithology, such as clay content, and geographic location, lead to different formation names, this is the case of the Meramec and Sycamore formations. The main difference between these formations is the clay content, being slightly higher in the Sycamore than the Meramec. Both formations were deposited as sediment-gravity flows in a mixed carbonate-siliciclastic system. Most of the public data came from the Meramec formation.

2. Methods

We use robust CNN architectures developed by computer vision specialists and previously put to test in a data-rich problem, thus we mainly focus on the adaptation of such CNN models to our domain-specific task: the microfacies classification problem.

Because grain size plays a crucial role in petrographic analysis, we use images with a consistent 10x magnification zoom. To compensate for the relatively low resolution of most CNN models used to construct the ILSVRC dataset (usually ranging between 200 by 200–400 by 400 pixels), we crop the original thin section photographs (1292 by 968 pixels) into a suite of smaller 644 by 644 pixels, overlapping square images (sub-images, Fig. 1) thereby augmenting the number of training images. Data augmentation increases the diversity of training samples thereby reducing overfitting (Cireşan et al., 2011; Takahashi et al., 2018). We eliminate the bottom right cropped images because many of them contain an alphanumeric scale bar (Fig. 1). The smaller images have enough resolution to be used for transfer learning, overlap between the sub-images helps to show that grain position is not important, and image size is sufficiently large to avoid isolating spurious bigger grains that could negatively impact the training.

Table 1
Original data used in this study. The thin sections are from the Mississippian Strata in the Ardmore basin, Oklahoma.

Microfacies	Number of thin sections	Number of photographs
Argillaceous siltstone	16	48
Bioturbated siltstone	29	87
Massive calcareous siltstone	15	45
Massive calcite-cemented siltstone	25	75
Porous calcareous siltstone	13	39

Table 2
Public data used as final test for this study.

Microfacies	Number of thin sections
Argillaceous siltstone	0
Bioturbated siltstone	25
Massive calcareous siltstone	18
Massive calcite-cemented siltstone	19
Porous calcareous siltstone	19
Lithofacies not present in training data (referenced as “Unknown”)	19

The image cropping process also increases the reliability in our final test data evaluation. Similar to how a petrographer classifies a thin section (or photograph of a thin section) based on an average of the visual aspect of the grains in the complete sample being analyzed, our model provides the classification based on the arguments of the maxima of the smaller sub-images. We call such an approach “voting” as the photograph of the thin section will be classified based on the microfacies with the most numbers of “votes”. Therefore, if a thin section image has most of its smaller sub-images labeled as argillaceous siltstone, the final lithofacies assigned by our model will be argillaceous siltstone. In cases in which there is not a single absolute maximum, we declare the model assigned a “tie” for the thin section image.

During initial training, we observed that most of the incorrect CNN prediction labeling was due to a poor color balance in the photographs within the same microfacies, with some images having a color shift to red or yellow. Such color shift occurs due to the difference in color temperature when light passes through the thin section and it goes through the objective lens. Bianco et al. (2017) studied the effects of color balancing and found that suitable color balancing yields a significant improvement in the accuracy for many CNN architectures. We follow Limare et al.’s (2011) methodology and compensate for the color shift assuming that the highest values of red, green, and blue observed in a photograph correspond to white, and the lowest values to black. Fig. 2 shows the effect of color balancing on a representative thin section.

After color balancing each image, we subdivide our thin section data from Table 1 into training, validation, and test data sets. The photographs of the thin section are ensured to remain in a single set, i.e., all sub-images of a photograph are either in the training, or validation, or test set, never in more than one set. The training set goes through another simple step of data augmentation in which we simply rotate the sub-images in 90, 180, and 270°; then we flip the initial smaller cropped image around the horizontal axis and rotate it 90, 180, and 270° again. Unlike other computer vision tasks in which the orientation or the relative position of an element is important for the overall performance, position and rotation of grains in a thin section are irrelevant. Table 3 shows the training, validation, and testing data set count after the preprocessing steps. These datasets are based on the sub-images and are available to download (supplementary materials) along with the original parallel polarized thin section photographs.

With the data prepared, we fine-tune four off-the-shelf pretrained different CNN models: VGG19 (Simonyan and Zisserman, 2014), MobileNetV2 (Sandler et al., 2018), InceptionV3 (Szegedy et al., 2015), and ResNet50 (He et al., 2016). More details of the models’ architectures are provided in their original references, here we provide a short summary of each one of them. VGG models are composed only of

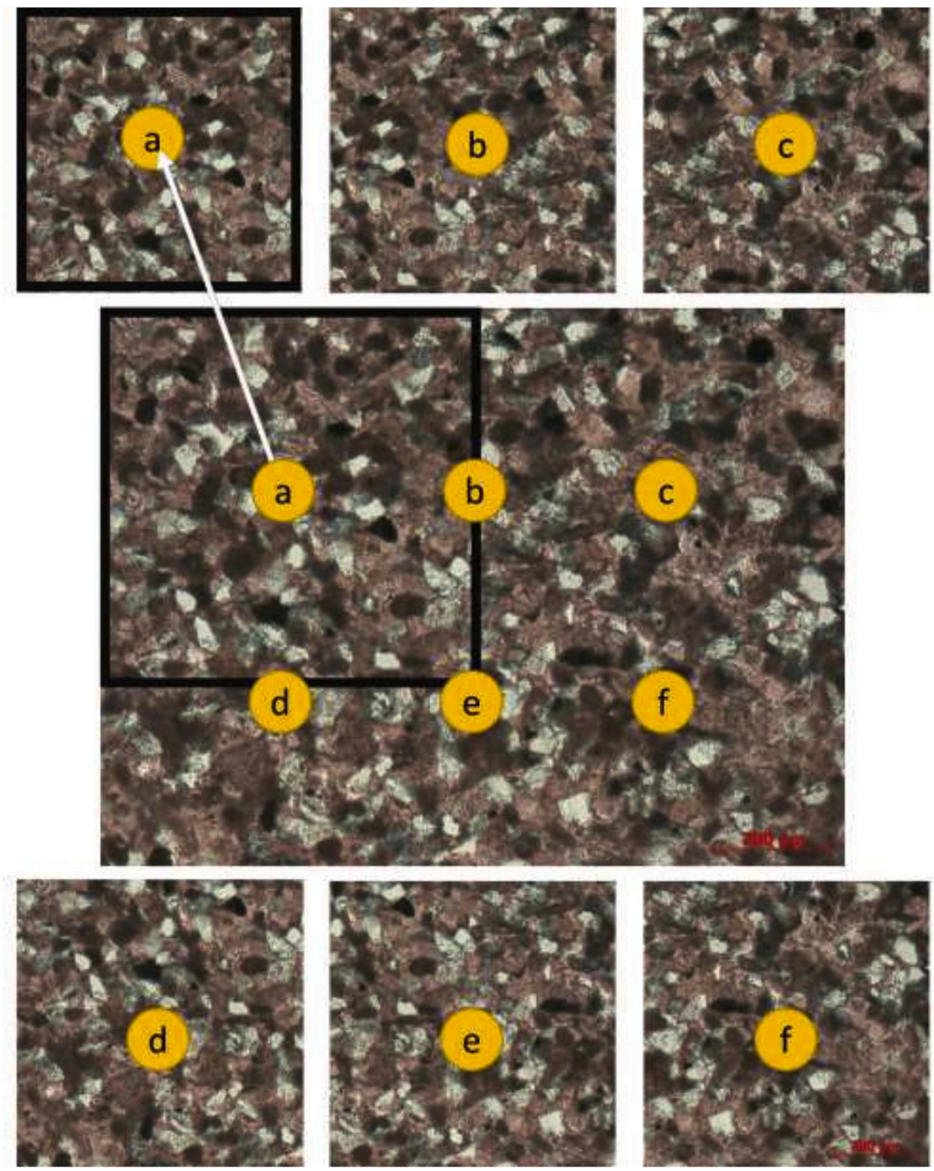

Fig. 1. An original photograph of a massive calcareous siltstone thin section (center, bigger) taken with 10x objective magnification and the sub-images used for training and testing (top and bottom rows, smaller). Sub-image a indicates with a black outline the boundaries and the center of the cropped image with a golden circle and the respective letter. The other sub-images are only represented by their center letters. Sub-image f is discarded in the training and validation set, as some original photographs will be marked with a scale bar. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

convolutional, pooling layers, and fully connected layers, and are therefore relatively simple when compared to the other architectures, and can be useful as a baseline. InceptionV3 is an improvement on GoogLeNet (Szegedy et al., 2014) and uses Inception blocks that combine convolutional filters with different sizes, as well as concatenation of filter outputs to achieve better performance. ResNets apply residual blocks, adding the output from the previous layer to the upcoming layer. The use of residual blocks made it possible to train deeper networks and helps with vanishing and exploding gradients. Finally, MobileNetV2 is an improvement on MobileNet (Howard et al., 2017) and makes use of depthwise separable convolution as building blocks. These architectures were chosen due to their popularity and ease of access to trained models. All models have convolutional layers on their base (i.e., closer to the input of the model) and terminate on fully connected layers (i.e., closer to the output of the model). Hereinafter we generically refer to "base model" as any of the pre-trained architectures described above; "top model" as the fully connected layers; and "convolutional layers" as the initial blocks, whether composed of Inception, residual, or other blocks. To perform fine-tuning, we discard the original top model and train a new classification network (or new top model) on top of the original convolutional layers of the base models.

The fine-tuning technique here is very similar to the one implemented by Yin et al. (2017):

1. Remove the top layers of the CNN model with ILSVRC parameters, and use the CNN model as the base model fixed feature extractor (traditional transfer learning, Yin et al., 2017). For all the base models, we maintain all convolutional layers to perform feature extraction. With the features extracted by the convolutional layers, we train a new classification network with five outputs (according to our number of classes/microfacies) by using Stochastic Gradient Descent (SGD) optimization. The new top model is simple, composed of a single dropout layer followed by a fully connected layer.
2. Combine the newly trained small classification network on the top of the base CNN model. We again use SGD with a small learning rate (le-4, reducing by a factor of 10 on plateaus), to update the parameters for the complete CNN model.

In other words, fine-tuning is a two-step process. In the first step, the new classification model initialized with random weights is trained using as input the features extracted by the convolutional layers of the pretrained CNN model (the base model). Thus, the convolutional filters

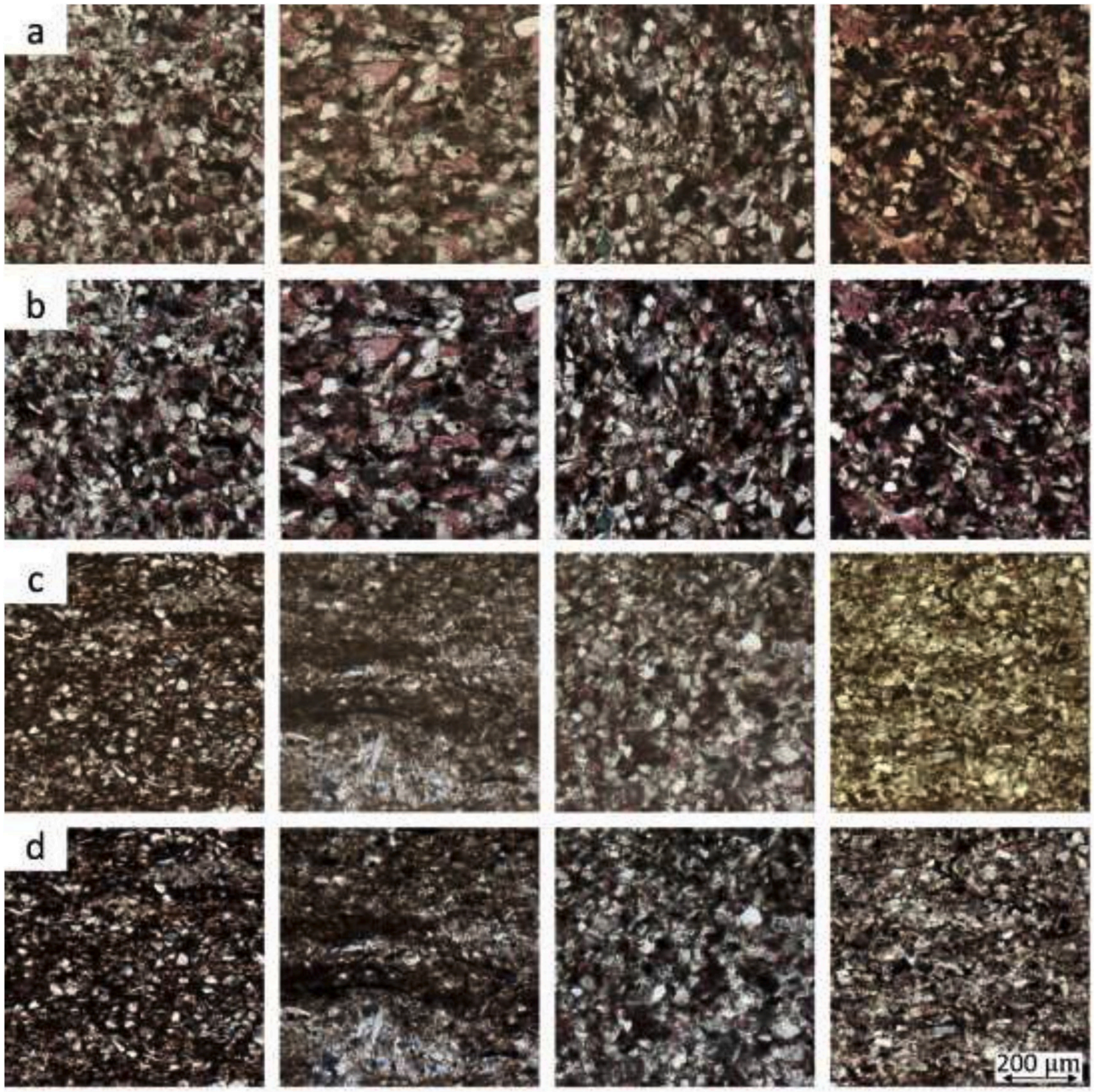

Fig. 2. Effects of color balancing. Row (a) examples of cropped photographs of massive calcareous siltstone before and row (b) after color balancing. Row (c) bioturbated siltstone before and (d) after color balancing. Note the examples in the last column. Sometimes photographs tend to be yellow, red or blue. The color balancing process helps to merge these images with the rest of the dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 3
Original data separated in training, validation, and test sets.

Lithofacies	Training set	Validation set	Test set
Argillaceous siltstone	880	55	90
Bioturbated siltstone	1200	110	190
Massive calcareous siltstone	680	70	80
Massive calcite-cemented siltstone	1160	120	125
Porous calcareous siltstone	640	30	85

of the base model are not updated. In the second step, the base and the new top model are combined, and the weights of the complete network are updated during training.

We use cross-entropy $H(\boldsymbol{p},\boldsymbol{q})$ during training:

$$H(\boldsymbol{p},\boldsymbol{q}) = -\sum_{c=1}^{C} \boldsymbol{p}_c \log(\boldsymbol{q}_c) \tag{1}$$

where C is the number of classes, logis the natural logarithm, $\boldsymbol{p}$ represents the true labels, and $\boldsymbol{q}$ the output of the last classification layer in the network. $H(\boldsymbol{p},\boldsymbol{q})$ represents the cost of a single sample and we minimize the loss, sum of costs of all samples, over all training samples. When we minimize the cross-entropy, we incentivize the CNN to increase the probability that the analyzed image to be assigned to the class c, when the image true label belongs to the class c.

We evaluate the performance of the fine-tuned models based on the test data separated from our original data set. We then select the best model and perform a final evaluation based on the classification our model provides to the public data. To perform the final evaluation, we use the six sub-images (Fig. 2) and three extra randomly centered sub-images with the same dimensions as the sub-images as shown in Fig. 2. These three extra sub-images help in the voting process to reduce the chances of ties.

3. Results

Table 4 shows the training, validation, and test set accuracy of the four fine-tuned CNN models, as well as the test set accuracy for the resulting thin section photograph voting. Table 4 also provides test set kappa metrics for the thin section photograph voting. We trained the models using a laptop with an NVIDIA GeForce GTX 1050 graphic

Table 4
Accuracy of sub-images, and accuracy and kappa for thin section photographs provided by fine-tuned models. The thin section receives the label according to the winning vote of its labeled smaller image crops.

Fine-tuned model	Accuracy				Kappa
	Training (sub-images)	Validation (sub-images)	Test (sub-images)	Test (photograph voting)	Test (photograph voting)
VGG19	1.00	0.93	0.93	0.95	0.93
MobileNetV2	1.00	0.90	0.91	0.94	0.92
InceptionV3	1.00	0.90	0.91	0.96	0.95
ResNet50	1.00	0.89	0.91	0.96	0.95

processing unit. Training is relatively fast, and the time to train one model ranged between one and 2 h. All models reach accuracies higher than 90% on the test set, however some overfitting is present as training data shows 100% accuracy. The use of augmentation improves accuracy by an average of 9 percentage points (pp) for the validation set, and 6pp for the test set. Results in Table 4 shows that the technique of

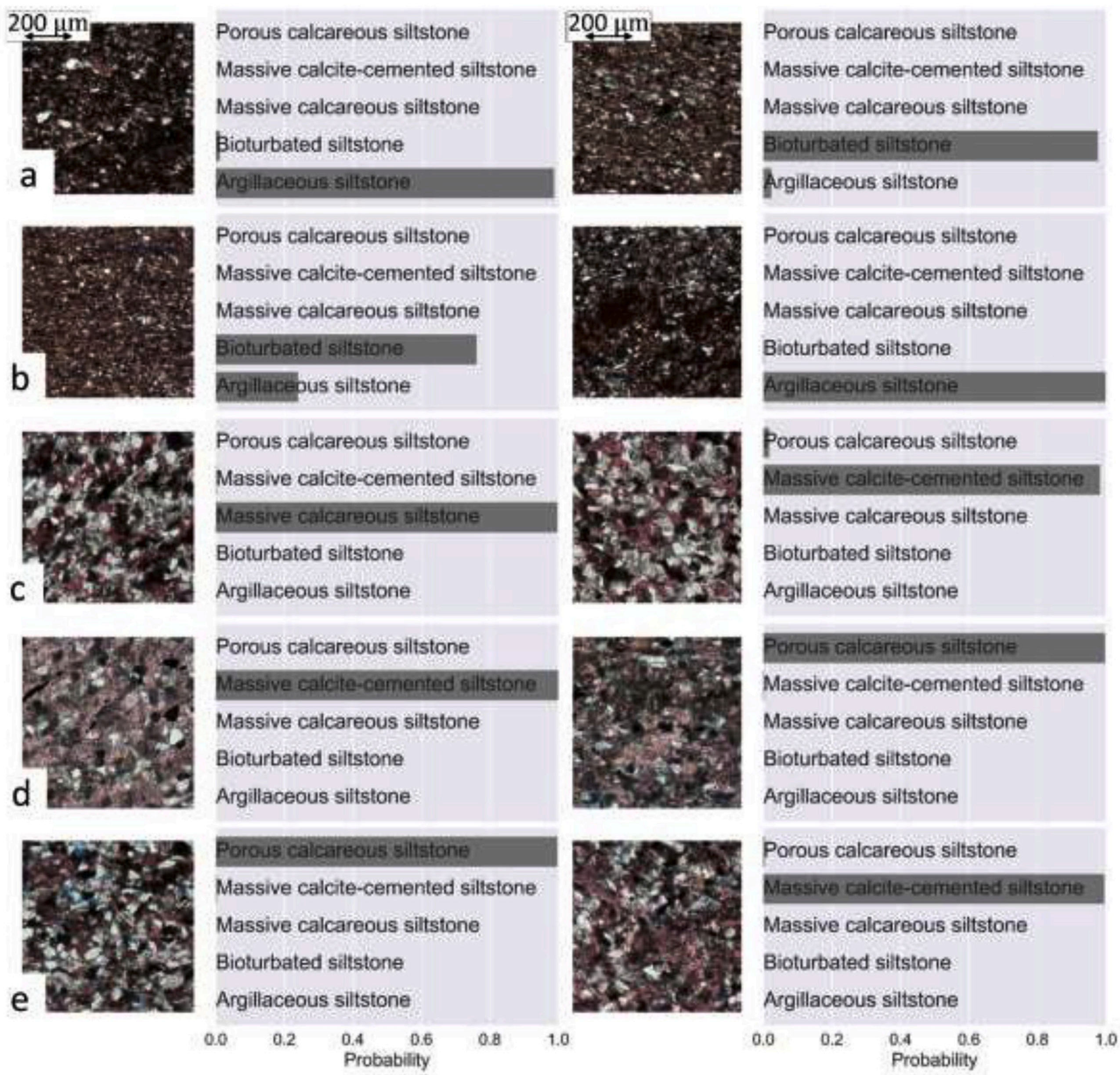

Fig. 3. Examples of classification provided by fine-tuned ResNet50 for the smaller cropped images in the test set. Images in the same row were extracted from the same microfacies as labeled by the interpreter. The left column shows examples of smaller cropped images in which the classification provided by the CNN model is the same as the classification provided by the petrographer. In contrast, the right column shows examples of smaller cropped images in which the classification provided by the CNN is not the same as the classification provided by the petrographer. Row (a) shows sub-images extracted from a photograph classified as argillaceous siltstone by the petrographer, row (b) was classified as bioturbated siltstone, (c) as massive calcareous siltstone, (d) massive calcite-cemented siltstone, and (e) porous calcareous siltstone.

photograph voting improves accuracy. Following McHugh (2012) for Cohen's kappa interpretation, all models achieve almost perfect (>0.90) level of agreement with the labels provided for the thin section photographs by the petrographer. The fine-tuned InceptionV3 and ResNet50 tied with the best accuracy (0.96) in the test set for photograph voting, but the overall difference in performance is small, showing that even the relatively simple VGG19 can properly classify most of the data.

Fig 3 shows examples of the resulting classification assigned by the fine-tuned ResNet50 to different sub-images of each one of the five classes present in the training data. Such examples provide a useful sample on the details of the prediction provided by the fine-tuned models. For each one of the five classes, we select thin section photographs of sub-images in which the fine-tuned ResNet50 assigned the same classification as the petrographer. Thus, Fig. 3 shows examples in which the class provided by the fine-tuned ResNet50 agrees with the classification provided by the petrographer, as well as examples in which the classification is different.

Fig. 4 shows the performance of the fine-tuned models compared against the petrographer-provided classification for the thin section photographs making use of confusion matrices. The results show that the majority of disagreements happen in microfacies containing very similar characteristics between each other, e.g., between argillaceous siltstone and bioturbated siltstone, and between massive calcite siltstone and massive calcite-cemented siltstone. MobileNetV2 (Fig. 4) is the only fine-tuned model that confuses massive calcite siltstone with porous calcareous siltstone. Such confusion might be indicative that in this experiment MobileNetV2 is slightly less sensitive to the blue color of the epoxy used to highlight porosity, as that is one of the main differences between massive calcite siltstone and porous calcareous siltstone. Yet, it is striking how minor differences between argillaceous siltstone and bioturbated siltstone are so well captured by the models. Bioturbation structures are minor lineaments or disturbances on the organization of grains caused by organisms, are often very subtle, and the only difference between argillaceous siltstone and bioturbated siltstone. Thus, it is encouraging CNN models are capable of using that information to differentiate between classes.

Lastly, we use the fine-tuned models to classify public data from the OPIC (Fig. 5). This evaluation of our model using public data serves as an initial evaluation of a possible multi-formation or multi-basin thin section CNN classifier. Results show a drastic decrease in the performance of the models, with accuracies below 50%. Metrics computed without consideration of the unknown microfacies (Table 5) are slightly better,

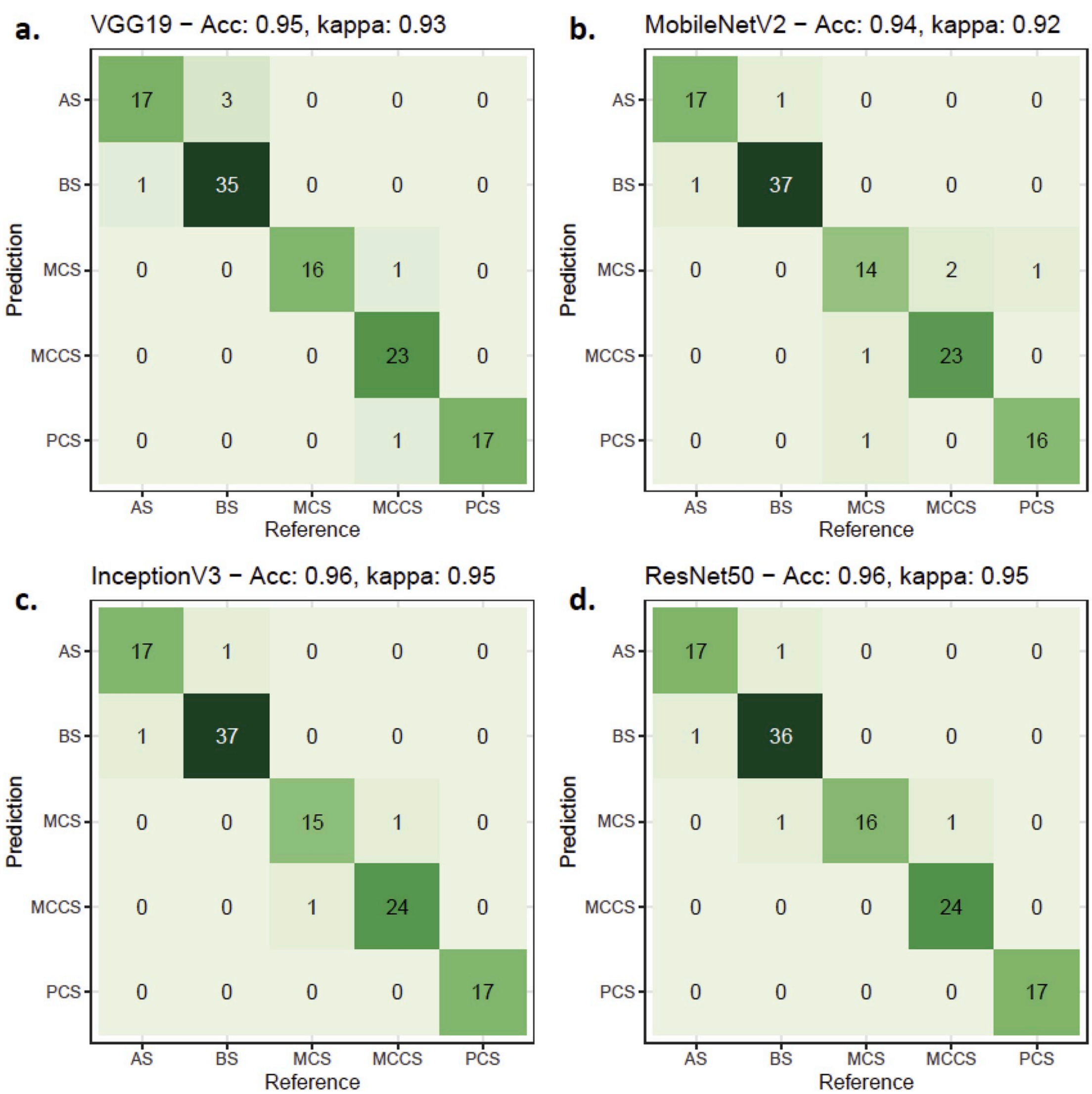

Fig. 4. Confusion matrix comparing the classification provided by the petrographer (reference) and the classification obtained with the fine-tuned models (prediction) for the test set thin section photographs. (a) shows the classification provided by the fine-tuned VGG19, as well as the accuracy and kappa values. (b), (c), and (d) show results for MobileNetV2, InceptionV3, and ResNet50 respectively. The class names are abbreviated: Argillaceous siltstone (AS), Bioturbated siltstone (BS), Massive calcareous siltstone (MCS), Massive calcite-cemented siltstone (MCCS), and Porous calcareous siltstones (PCS).

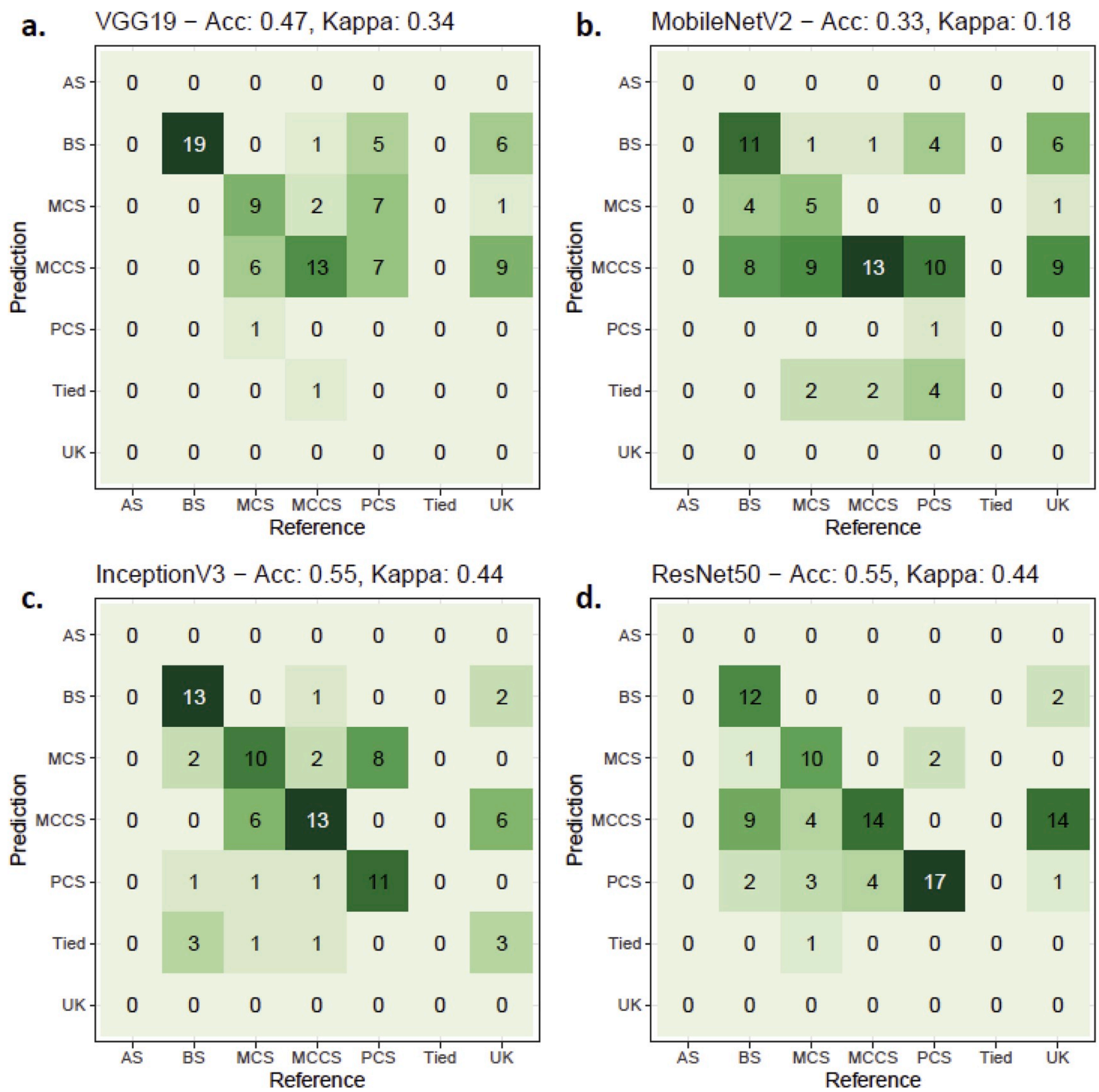

Fig. 5. Confusion matrix comparing the classification provided by the petrographer (reference) and the classification obtained with the fine-tuned models (prediction) for the final public data test set thin section photographs. (a) shows the classification provided by fine-tuned VGG19, as well as the accuracy and kappa values. (b), (c), and (d) show results for MobileNetV2, InceptionV3, and ResNet50 respectively. The class names are abbreviated: Argillaceous siltstone (AS), Bioturbated siltstone (BS), Massive calcareous siltstone (MCS), Massive calcite-cemented siltstone (MCCS), and Porous calcareous siltstones (PCS).

with a kappa indicating minimal agreement for MobileNetV2 and VGG, and weak agreement for other models. Moreover, unlike the results obtained with the original data (Table 4 and Fig. 4), the disagreement occurs for microfacies that are significantly different. For example, MobileNetV2 classifies photographs originally classified as bioturbated siltstone as massive calcite-cemented siltstone eight times (shown in Fig. 6). Despite being the same microfacies, the samples of bioturbated siltstone in the training set are different than the ones in the public data. Although the voting process helps classification, the strategy to extract three randomly centered sub-images, adding to those sub-images shown in Fig. 1, is not very robust in this case and accuracy can vary significantly with different random sub-images. In the experiments with the worst performance, the accuracy ignoring unknowns for the best performing ResNet50 and InceptionV3 were low as 47% and 43% respectively. Similar variation in performance due to different random sub-images is not observed with the original data, indicating models are more robust when the test set has similar characteristics to the training set. As we continue to add more training data and better adapt our CNN models, we anticipate further acceleration and accuracy of thin section analysis.

Table 5
Accuracy and kappa for thin section photographs provided by fine-tuned models for the public data. The metrics in this table are computed ignoring the "unknown" samples.

Fine-tuned model	Accuracy	Kappa
VGG19	0.51	0.37
MobileNetV2	0.38	0.21
InceptionV3	0.59	0.48
ResNet50	0.66	0.56

4. Discussion

To the authors' knowledge, this is one of the few studies for automated microfacies classification with CNN using rock thin sections. In the methodology we implement, a user can take multiple photographs of a single thin section, and obtain its classification as predicted by the model. Based on our tests, the accuracy of the procedure presented here is comparable to accuracies of a petrographer, as long as the lithofacies being analyzed were present in the training data and the thin sections were processed with similar methodology. Our study is different than

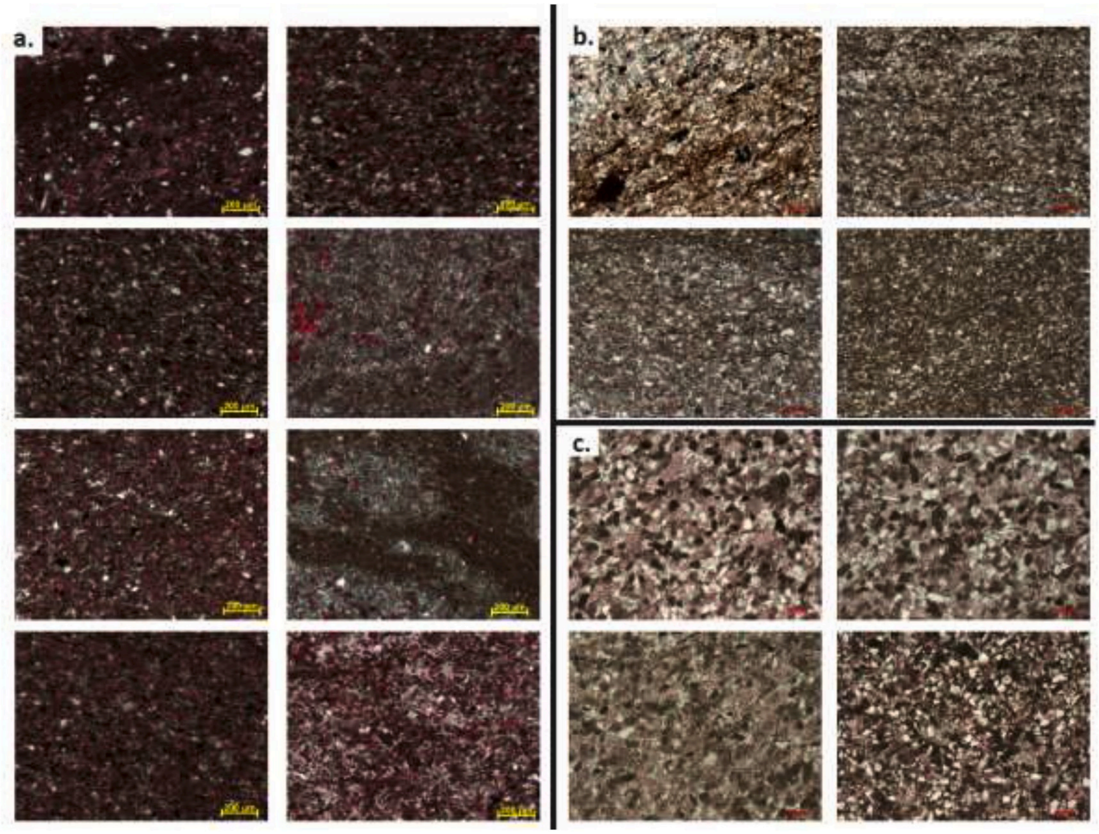

Fig. 6. Examples of the eight misclassified bioturbated siltstone thin sections from the public data and samples from training set. (a) the eight bioturbated siltstones thin sections classified as massive calcite-cemented siltstones by MobileNetV2. (b) examples of bioturbated siltstones in the training set (images that the CNN models evaluated during training). (c) examples of massive calcite-cemented siltstones in the training set.

that of Cheng and Guo (2017) because we differentiate between five different microfacies, whereas Cheng and Guo (2017) differentiate between three granulometric classifications. In some sense, the analysis performed by Budennyy et al. (2017) using thin sections is more complete than the one we provide in this study. Budennyy et al. (2017) were able to classify the mineral composition of sandstones with an accuracy of 80%. Their technique relies on watershed segmentation methods to isolate mineral grains and extract features before further analysis. Such an approach is interesting in the fact that it generates data useful to perform a more complete analysis of the thin section (e.g., making possible to analyze roundness, grain size, and others), but also introduces another step that needs to be quality controlled by domain experts. One of the disadvantages of classical feature extraction methodologies is their heavy dependence on human intervention. Moreover, the choice of features to be used for analysis is time consuming and frequently depends on heuristic design decisions. Neither Cheng and Guo (2017), Budennyy et al. (2017), or Karimpouli and Tahmasebi (2019) provided examples of the performance of their model when tested with significantly different data, as we present in our public data evaluation.

Unlike a human interpreter who relies upon a defined set of morphological measurements to perform microfacies classifications, the CNN operates from no knowledge of specific attribute analysis and performs the classification based on image characteristics. CNN labeled datasets have the potential to reduce petrographer bias, yielding a reduced inconsistency on thin sections classification. When analyzing a new image, the CNN model (as implemented in this study) will always generate a set of probabilities that such image belongs to the CNN's learned microfacies. For that reason, Fig. 5 shows that the CNN provides classifications for all the thin sections classified as unknown by the petrographer. The number of unknowns can be reduced when more examples of microfacies are provided to the CNN models.

Fig. 4 indicates that the CNN misclassifications are in fact similar to the description a petrographer would assign to a particular section of a thin section photograph. Due to thin section heterogeneities, the CNN classification maybe is may be correct for the particular sub-image in analysis, but accuracy generally increases when multiple sub-images are used. Therefore, our voting scheme then is helpful as it reduces possible misconceptions. One of the explanations for the misclassification is the criteria that the petrographer used for thin section microfacies classification. There are two main groups of rock types: structureless or massive, and structured. To divide the microfacies within these two main groups, the petrographers use a qualitative-visual criterion. For example, the massive siltstones can be calcareous, porous, and calcite-cemented. However, the criteria used to divide between them were the visual content of calcite cement and porosity and no statistical method was used to quantify the proportion of cement or porosity. Thus, the misclassifications in the original dataset (Table 4 and Fig. 5), are mostly caused due to the fact that the microfacies are classified based on gradual boundaries and the division between classes can be somewhat subjective. We suggest including other data to quantify the amount of cement, mineralogy and porosity. With a more quantitative interpretation, we can reduce the interpretation bias.

Fig. 5 and Table 5 show that the performance is greatly reduced when models are used to classify public data. The misclassifications in the public data are mostly due to different staining, as well as different mineralogy composition and imaging, changing the image characteristics. The models were built with thin sections stained with Alizarin red for calcite identification, and blue epoxy for porosity identification. However, public data thin sections do not always have these features. Therefore, thin sections with high calcite content could be labeled as microfacies without calcite. In fact, most of the confusion between massive calcite-cemented siltstones and the calcareous siltstones could be explained by the lack of alizarin stains.

Finally, the photograph by itself plays an important role in the model and so can contribute to the label bias. The original labels resulted from the observation of the actual thin sections under the microscope and not based on the photographs. Dozens of different photographs without any overlap can be taken from the same thin section with 10X objective magnification. The photographs we captured for this study were taken randomly in different locations of the thin section. What differentiates between argillaceous and bioturbated siltstones are the bioturbation patterns. Bioturbation is evident when the thin sections are examined under the microscope, however, sometimes the bioturbation evidence is obscured when cropping the thin section images into smaller 10X photographs. Thus, to avoid misclassification the photographs should depict the criteria used by the petrographer for the original classification. This difficulty in capturing complete characteristics of the entirety of the thin sections with random photographs indicates that most of the misclassification is the result of the preparation and labeling of the data used to train the model rather than the CNN model by itself. This misclassification pattern also shows a potential improvement that the use of CNN models can provide. If the thin section is captured in its entirety, the CNN can quickly provide classifications for all its sections. A petrographer can then quality control the CNN results as well as easily note outliers that could either be mistakes or important features that can be further analyzed.

As the digitization of legacy data accelerates, and thin section preparation and data storage methodologies are standardized, the approach presented here can improve with more detailed and directed image processing. Image segmentation techniques can be used to differentiate between different minerals, which can be a powerful tool for microfacies classification. The technique we demonstrate in this manuscript is very general and can easily be modified to suit the identification of thin sections coming from different formations.

5. Conclusions

In this paper, we propose the use of transfer learning and fine-tuning of robust CNN models for petrographic thin section classification, achieving accuracies above 90% for all the models tested when using data from the Sycamore Formation and obtained with the same type of processing. Furthermore, using public data, we investigate how such fine-tuned CNN models can be used to classify sediment-gravity flows in a mixed carbonate-siliciclastic systems from the Sycamore and Meramec formations with significantly different parameters; however, current results show a drastic reduction in the model's performance. It is likely fine-tuning could perform well using thin section photographs of sediments in other depositional settings or other basins, as long as enough data is available.

We focus on the use of parallel polarized petrographic thin section images, as they are sufficient to differentiate between the classes/microfacies present in our dataset. Cross-polarized images could be included for the cases in which such imaging technique is crucial for proper lithofacies classification, for example to differentiate between a rock enriched in quartz grains and a rock enriched in feldspars grains. In addition, this paper mostly concentrates on the use of CNN models at a specific 10x magnification level. As different lithological and diagenetic properties can only be analyzed in different scales, many other studies can be conducted with a similar technique. We believe that the implementation of the methodology we discuss here has the potential to further improve petrographic thin section classification speed and help geoscientists make use of such invaluable data.

Authorship statement

RPL developed the conception and design of study, wrote the necessary scripts, performed analysis, and wrote the manuscript. DD acquired the data, participated in the analysis, and helped writing the manuscript. CN guided the analysis, helped writing the manuscript, and revised the manuscript critically for important intellectual content. RS helped writing the manuscript, and revised the manuscript critically for important intellectual content. KJM helped in the conception and design of study, helped writing the manuscript, and revised the manuscript critically for important intellectual content.

Declaration of competing interest

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

CRediT authorship contribution statement

Rafael Pires de Lima: Conceptualization, Software, Formal analysis, Writing - original draft. **David Duarte:** Data curation, Formal analysis, Writing - original draft. **Charles Nicholson:** Formal analysis, Writing - review & editing. **Roger Slatt:** Funding acquisition, Writing - review & editing. **Kurt J. Marfurt:** Conceptualization, Writing - review & editing.

Acknowledgments

Pires de Lima acknowledges Brazilian National Council for Scientific and Technological Development (CNPq) for the financial support (grant 203589/2014-9) and CPRM for granting the leave of absence allowing the pursuit of his Ph.D. studies. The acquisition of the data used to build the CNN model was funded through University of Oklahoma's Institute of Reservoir Characterization Consortium. We thank the reviewers that helped us improve the quality of the manuscript. We thank Katie Welch who helped review the manuscript. We would like to acknowledge our mentor, colleague, and co-author Dr. Roger Slatt who unfortunately passed away before the final revision of this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cageo.2020.104481.

References

Bengio, Y., 2012. Deep learning of representations for unsupervised and transfer learning. In: Guyon, I., Dror, G., Lemaire, V., Taylor, G., Silver, D. (Eds.), Proceedings of ICML Workshop on Unsupervised and Transfer Learning, Proceedings of Machine Learning Research. PMLR, Bellevue, Washington, USA, pp. 17–36.

Bianco, S., Cusano, C., Napoletano, P., Schettini, R., Bianco, S., Cusano, C., Napoletano, P., Schettini, R., 2017. Improving CNN-based texture classification by color balancing. J. Imaging 3, 33. https://doi.org/10.3390/jimaging3030033.

Budennyy, S., Pachezhertsev, A., Bukharev, A., Erofeev, A., Mitrushkin, D., Belozerov, B., 2017. Image processing and machine learning approaches for petrographic thin section analysis. SPE Russ. Pet. Technol. Conf. https://doi.org/10.2118/187885-MS.

Carranza-Rojas, J., Goeau, H., Bonnet, P., Mata-Montero, E., Joly, A., 2017. Going deeper in the automated identification of Herbarium specimens. BMC Evol. Biol. 17, 181. https://doi.org/10.1186/s12862-017-1014-z.

Caruana, R., 1995. Learning many related tasks at the same time with backpropagation. In: Tesauro, G., Touretzky, D.S., Leen, T.K. (Eds.), Advances in Neural Information Processing Systems 7. MIT Press, pp. 657–664.

Chen, Z., Zhang, T., Ouyang, C., Chen, Z., Zhang, T., Ouyang, C., 2018. End-to-End airplane detection using transfer learning in remote sensing images. Rem. Sens. 10, 139. https://doi.org/10.3390/rs10010139.

Cheng, G., Guo, W., 2017. Rock images classification by using deep convolution neural network. J. Phys. Conf. Ser. 887, 012089 https://doi.org/10.1088/1742-6596/887/1/012089.

Cheng, G., Yue, Q., Qiang, X., 2018. Research on feasibility of convolution neural networks for rock thin sections image retrieval. In: 2018 2nd IEEE Adv. Inf. Manag. Autom. Control Conf, pp. 2539–2542.

Chollet, F., 2016. Xception: Deep Learning with Depthwise Separable Convolutions. CoRR abs/1610.0.

Cireşan, D.C., Meier, U., Masci, J., Gambardella, L.M., Schmidhuber, J., 2011. Flexible, high performance convolutional neural networks for image classification. In: Proceedings of the Twenty-Second International Joint Conference on Artificial

Intelligence - Volume Volume Two, IJCAI'11. AAAI Press, pp. 1237–1242. https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-210.
Cunha, A., Pochet, A., Lopes, H., Gattass, M., 2020. Seismic fault detection in real data using transfer learning from a convolutional neural network pre-trained with synthetic seismic data. Comput. Geosci. 135, 104344. https://doi.org/10.1016/j.cageo.2019.104344.
Datta, R., Joshi, D., Li, J., Wang, J.Z., 2008. Image retrieval: ideas, influences, and trends of the new age. ACM Comput. Surv. 40.
Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118. https://doi.org/10.1038/nature21056.
Gomez Villa, A., Salazar, A., Vargas, F., 2017. Towards automatic wild animal monitoring: identification of animal species in camera-trap images using very deep convolutional neural networks. Ecol. Inf. 41, 24–32. https://doi.org/10.1016/J.ECOINF.2017.07.004.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.
Hong, Y., Ling, C., Ye, Z., 2018. End-to-end soccer video scene and event classification with deep transfer learning. In: 2018 International Conference on Intelligent Systems and Computer Vision (ISCV). IEEE, pp. 1–4. https://doi.org/10.1109/ISACV.2018.8369043.
Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications arXiv e-prints arXiv:1704.04861.
Huang, G., Liu, Z., Weinberger, K.Q., 2016a. Densely Connected Convolutional Networks. CoRR abs/1608.0.
Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K., 2016b. Deep Networks with Stochastic Depth.
Karimpouli, S., Tahmasebi, P., 2019. Segmentation of digital rock images using deep convolutional autoencoder networks. Comput. Geosci. 126, 142–150. https://doi.org/10.1016/J.CAGEO.2019.02.003.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems, vol. 1. Curran Associates Inc., USA, pp. 1097–1105. NIPS'12.
Launeau, P., Robin, P.-Y.F., 1996. Fabric analysis using the intercept method. Tectonophysics 267, 91–119. https://doi.org/10.1016/S0040-1951(96)00091-1.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539.
Limare, N., Lisani, J.-L., Morel, J.-M., Petro, A.B., Sbert, C., 2011. Simplest color balance. Image Process. Line 1, 297–315. https://doi.org/10.5201/ipol.2011.llmps-scb.
Machine Learning Glossary | Google Developers [WWW Document], URL https://developers.google.com/machine-learning/glossary/#top_of_page (accessed 1.21.19).
Maitre, J., Bouchard, K., Bédard, L.P., 2019. Mineral grains recognition using computer vision and machine learning. Comput. Geosci. 130, 84–93. https://doi.org/10.1016/j.cageo.2019.05.009.
McHugh, M.L., 2012. Interrater reliability: the kappa statistic. Biochem. medica 22, 276–282.
Mou, L., Ghamisi, P., Zhu, X.X., 2017. Deep recurrent neural networks for hyperspectral image classification. IEEE Trans. Geosci. Rem. Sens. 55, 3639–3655. https://doi.org/10.1109/TGRS.2016.2636241.
Nasseri, M.H.B., Mohanty, B., 2008. Fracture toughness anisotropy in granitic rocks. Int. J. Rock Mech. Min. Sci. 45, 167–193. https://doi.org/10.1016/J.IJRMMS.2007.04.005.
Olah, C., Mordvintsev, A., Schubert, L., 2017. Feature visualization. Distill. https://doi.org/10.23915/distill.00007.
Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., Mordvintsev, A., 2018. The Building Blocks of Interpretability. Distill. https://doi.org/10.23915/distill.00010.
Pires de Lima, R., Bonar, A., Coronado, D.D., Marfurt, K., Nicholson, C., 2019a. Deep convolutional neural networks as a geological image classification tool. Sediment. Rec. 17, 4–9. https://doi.org/10.2110/sedred.2019.2.4.
Pires de Lima, R., Suriamin, F., Marfurt, K.J., Pranter, M.J., 2019b. Convolutional neural networks as aid in core lithofacies classification. Interpretation 7, SF27–SF40. https://doi.org/10.1190/INT-2018-0245.1.
Přikryl, R., 2001. Some microstructural aspects of strength variation in rocks. Int. J. Rock Mech. Min. Sci. 38, 671–682. https://doi.org/10.1016/S1365-1609(01)00031-4.
Qayyum, A., Anwar, S.M., Awais, M., Majid, M., 2017. Medical image retrieval using deep convolutional neural network. Neurocomputing 266, 8–20. https://doi.org/10.1016/J.NEUCOM.2017.05.025.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252. https://doi.org/10.1007/s11263-015-0816-y.
Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C., 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks (ArXiv e-prints).
Simonyan, K., Vedaldi, A., Zisserman, A., 2013. Deep inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. CoRR abs/ vol. 1312, 6.
Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. ArXiv e–prints.
Sudakov, O., Burnaev, E., Koroteev, D., 2019. Driving digital rock towards machine learning: predicting permeability with gradient boosting and deep neural networks. Comput. Geosci. 127, 91–98. https://doi.org/10.1016/j.cageo.2019.02.002.
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2014. Going Deeper with Convolutions. CoRR abs/1409.4.
Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2015. Rethinking the Inception Architecture for Computer Vision. CoRR abs/1512.0.
Takahashi, R., Matsubara, T., Uehara, K., 2018. Data Augmentation Using Random Image Cropping and Patching for Deep CNNs. CoRR abs/1811.0.
Yang, X., Ye, Y., Li, X., Lau, R.Y.K., Zhang, X., Huang, X., 2018. Hyperspectral image classification with deep learning models. IEEE Trans. Geosci. Rem. Sens. 56, 5408–5423. https://doi.org/10.1109/TGRS.2018.2815613.
Yin, X., Chen, W., Wu, X., Yue, H., 2017. Fine-tuning and visualization of convolutional neural networks. In: 2017 12th IEEE Conference on Industrial Electronics and Applications (ICIEA). IEEE, pp. 1310–1315. https://doi.org/10.1109/ICIEA.2017.8283041.
Yosinski, J., Clune, J., Bengio, Y., Lipson, H., 2014. How transferable are features in deep neural networks? Adv. Neural Inf. Process. Syst. 27, 3320–3328.

Glossary

Accuracy: the fraction of total objects correctly classified. Values range from 0.0 to 1.0 (equivalently, 0%–100%). Accuracy equals to 1.0 means all classifications were correct, accuracy equals to 0.0 means all classifications were incorrect
Convolution: a mathematical operation that combines two functions producing an output. In machine learning applications, a convolutional layer uses two discrete functions, the input data and a convolutional kernel, to train the convolutional kernel weights
Convolutional Neural Networks (CNN): a neuron network architecture in which at least one layer is a convolutional layer
Deep Learning (DL): an artificial neural network architecture that contains more than one hidden layer
Fine-Tuning: the process of adjusting machine learning model parameters of a pre-trained model to improve performance for a specific problem type
Kappa: Cohen's kappa coefficient. A metric that takes into the consideration the agreement by chance. Kappa is given by $\kappa\frac{p_0-p_e}{1-p_e}$, where p_0 is the accuracy and p_e is the hypothetical probability of agreement by chance
Label: the names applied to an instance, sample, or example (for image classification, an image) associating it with a given class
Layer: a group of neurons in a machine learning model that process a set of input features
Machine Learning (ML): a collection of approaches in which systems improve their performance through automatic analysis of data
Neural Networks (NN): a machine learning model that combines linear and nonlinear transformations, loosely inspired in the behavior of brain neurons. It is typically organized in layers where each layer contains a number of nodes (or neurons)
Neuron: A node in a neural network, typically taking in multiple input values and generating one output value. The neuron calculates the output value by applying an activation function (nonlinear transformation) to a weighted sum of input values
Training: the process of finding the most appropriate weights of a machine learning model
Transfer Learning: a technique that uses information learned in a primary machine learning task to perform a secondary machine learning task
Top-X error: a measure of model accuracy. A classification is considered correct as long as the correct label is in one of the top X guessed labels. Top-1 error is the ratio of the incorrect classifications over the total number of classifications (1.0 minus accuracy)
Weights: the coefficients of a machine learning model. In a simple linear equation, the slope and intercept are the weights of the model. In CNNs, the weights are the convolutional kernel values. The training objective is to find the ideal weights of the machine learning model.This glossary presents common denominations in machine learning applications used throughout the manuscript. For a more comprehensive list, we refer the reader to Google's machine learning glossary ("Machine Learning Glossary | Google Developers," n.d.).