Supplement To:

Applications of Deep Learning to Decorated Ceramic Typology and Classification: A Case Study Using Tusayan White Ware from Northeast Arizona

Leszek Pawlowicz*, Christian E. Downum

Department of Anthropology, Northern Arizona University, PO Box 15200, Flagstaff, Arizona 86011-5200

1. Creating The Labeled Dataset For CNN Model Training

1.1 Typing and Analyzing the Original Set of Images

This study required model-based classifications of a large sample (thousands) of archaeological specimens. To achieve this, the authors chose Tusayan White Ware as an ideal candidate for digital classification. Sherds of Tusayan White Ware are abundant and readily accessible in regional museum collections, and hundreds of photographs and line drawings of sherds and vessels exist in publications. The clear contrast between the dark paint and the light background on painted sherds ensures that a high percentage of Tusayan White Ware specimens have an acceptable digital image for analysis. Further, with such a high degree of contrast, we could acquire images using a wide variety of digital cameras having varying levels of lens quality and resolution. The sherd images used came from the collections of the Museum of Northern Arizona and Northern Arizona University.

As expected, we experienced no significant problems with data consistency when we generated our own digital images taken with different cameras under a range of lighting conditions. Similarly, even legacy data, composed of published images of variable quality, worked well for this analysis. Initially, the sherd images were manually “cut out” of the original photographs with image editing software (Adobe Photoshop in this case), and pasted onto a white background with a square aspect ratio, the latter to conform to the requirements of most convolutional neural network models. Doing this manually was a tedious and time-consuming process; we are developing software that will automatically detect and crop sherd images from photographs taken on a dark background and save them in the appropriate format. This is an important factor to consider for future image-based classification projects because a potential barrier to success is the time required to accumulate the necessary large samples of artifact images. With an automated process of digital image detection and extraction, it is possible to include dozens of objects in a single digital photo. This maximizes the number of artifacts that can be photographed in the often-limited time available in archaeological collections repositories, and it minimizes the time and expense of processing the images.

In addition to the sherd images, contemporary archaeologists with expertise in the Tusayan White Ware typology were required to classify the sherds into types as a basis for training the CNN models. For this, we recruited four of the most experienced Tusayan White Ware analysts from the northern Arizona community of archaeological practice. Each of the four participants in this study (including the second author) has 30 or more years of field and laboratory analysis of ceramics, and each during their careers has personally classified tens of thousands of Tusayan White Ware potsherds. To encourage independent classifications without fear of being stigmatized for disagreement with others' type assignments or otherwise poor results, the senior author anonymized the classifiers' identities as analysts A, B, C, and D.

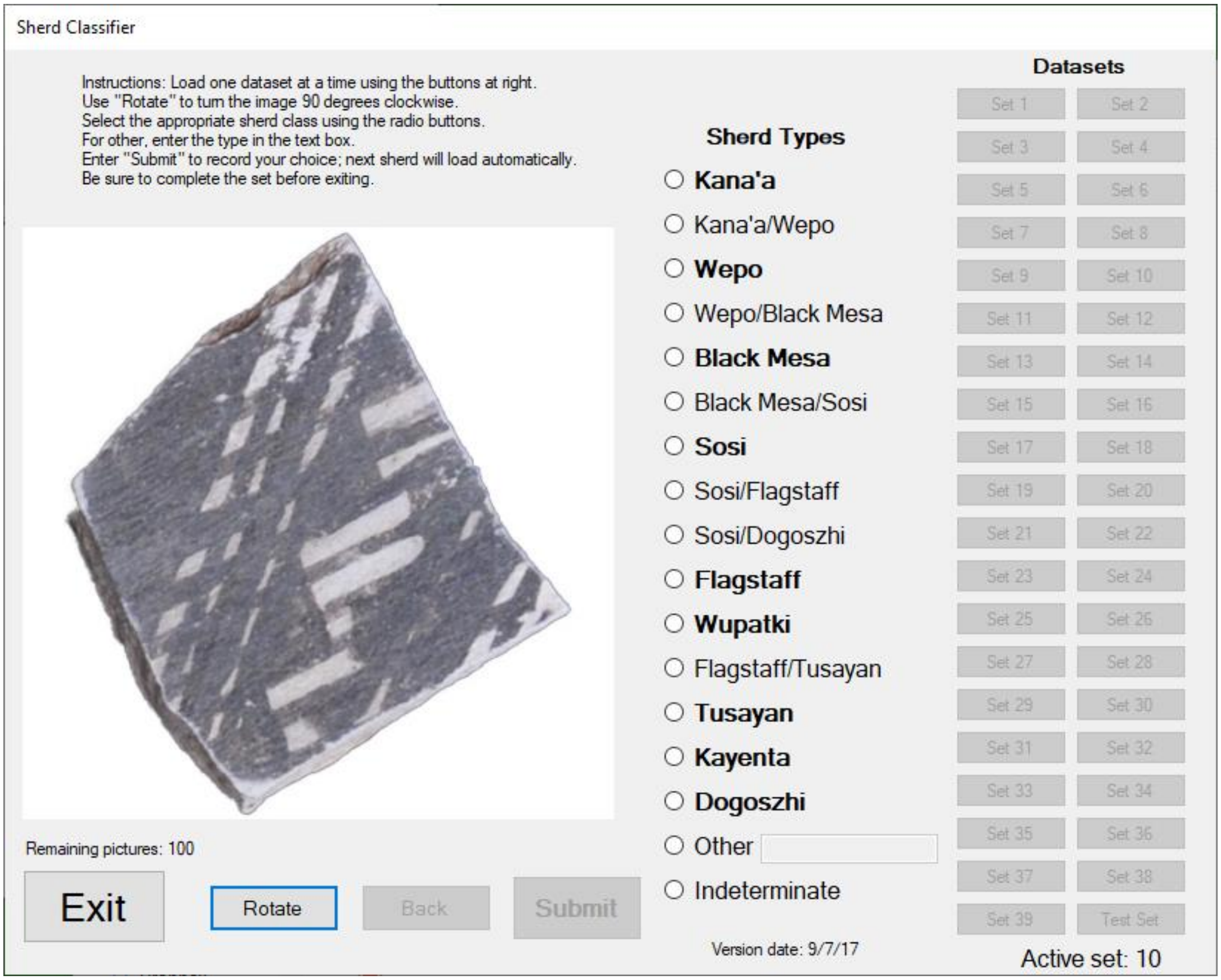
In order to train CNN models, a "labeled dataset" is needed, consisting of images assigned an accurate Tusayan White Ware type label by the expert analysts. We collected an initial set of approximately 3,250 Tusayan White Ware sherd photographs; of these, 3,064 were judged to have sufficient image quality, size, and design representation for accurate classification. To facilitate classification of digital images into pottery types, the senior author wrote a simple program to allow sherd analysts to classify sherds at their convenience. The program displays each sherd photograph and allows it to be rotated to vary the view. Analysts then classified individual sherds by selecting the proper type category from a radio-button menu (Figure S-1). Results were automatically written to database-compatible output files. The sherds were broken up into sets of 100, to allow the classifiers to parse their work into chunks that could be fit into their schedule, rather than having to classify 3,064 sherds all at once. The individual output files of 100, consisting of the name of the sherd image file and the assigned classification, were then combined and collated into a single spreadsheet containing all of the classifier results.

For each sherd, analysts had the option of selecting from the following common Tusayan White Ware types:

- Kana'a
- Wepo
- Black Mesa
- Sosi
- Dogoszhi
- Flagstaff
- Wupatki
- Tusayan
- Kayenta
- Other
- Indeterminate/Undifferentiated

In addition, they were given the option to select intermediate/mixed types:

- Kana'a /Wepo
- Wepo /Black Mesa
- Black Mesa / Sosi
- Sosi / Dogoszhi
- Sosi / Flagstaff
- Flagstaff / Tusayan



Instructions: Load one dataset at a time using the buttons at right.
Use "Rotate" to turn the image 90 degrees clockwise.
Select the appropriate sherd class using the radio buttons.
For other, enter the type in the text box.
Enter "Submit" to record your choice; next sherd will load automatically.
Be sure to complete the set before exiting.

Remaining pictures: 100

Version date: 9/7/17

Active set: 10

Figure S-1 – Program used to classify sherd images by Tusayan White Ware type.

These mixed types were included in the hope that they could be used to identify ceramics with designs transitional between different types. However, classifiers reported using mixed type categories mostly to reflect their uncertainty as to which type a sherd belonged, rather than using them to categorize sherds with mixed design elements that could be interpreted as transitional or intermediate examples of two types. Fortunately, these mixed types were used sparingly by the classifiers, and consequently did not affect the usability of the results. Future classification efforts will eliminate the option to choose intermediate types.

“Other” was mostly used to identify sherds by the human classifiers that were either very uncommon TWW types, or Little Colorado White Ware types; the latter ware contains types with design characteristics very similar to Tusayan White Ware. However, the latter identifications were erroneous, since all of the sherds had been physically pre-screened to ensure that they were Tusayan White Ware (sand temper, light paste, usually non-slipped), and not Little Colorado White Ware (sherd temper, dark paste, slipped). Indeterminate/Undifferentiated meant that the classifier did not feel confident enough to assign a type to a sherd.

1.2 Evaluating the Results From The Human Classifiers

Over a period of approximately 6 months, the four sherd analysts classified the 3,064 sherds and submitted their individual databases to the senior author. Results were totaled and correlated by assigned type for all four classifiers. For each sherd assigned to a single type only, a value of one was added to the total for that type. For split classifications, a value of ½ was added to the total. For example, a type designation of “Black Mesa” added one to the Black Mesa total, while a split type of “Black Mesa / Sosi” added ½ to the Black Mesa total, and ½ to the Sosi total. Table S-I lists total counts by type for each classifier.

Except for Dogoszhi, the variation in specific type totals is surprisingly high. Dogoszhi has a very distinctive style of hatching between delineated borders, so the relatively good results for this type are not surprising. In a previous study of the replicability of Tusayan White Ware sherd typing (Fish 1976, 1978), Dogoszhi was the only type for which all classifiers agreed when they assigned types to Tusayan White Ware and Gray Ware ceramics. Wepo is not as well-defined or commonly used as other types, so the high variation is also not surprising; one of the classifiers expressed a strong skepticism that Wepo was even a type that could be distinguished from its temporal/stylistic predecessor (Kana’a) and successor (Black Mesa). There was also a wide range of totals for “Indeterminate/Undifferentiated”, sherds that the classifiers decided did not have enough clear design characteristics to allow them to be typed accurately. The high variation in type counts immediately suggested that there would be substantial numbers of sherds where the classifiers would not agree on the type.

The data also showed that Wupatki was a problematic type. Only two of the classifiers identified more than 1% of the sherds as Wupatki (Table S-I), and only one sherd was typed by any two of the classifiers as Wupatki. Based on these results, it seems that Wupatki, if it exists as a legitimate Tusayan White Ware type, does not have a well-defined set of distinguishing characteristics applied consistently by the current archaeological community. Consequently, it was dropped from this study as a type.

Classification results were calculated based on the agreement between individual pairs of classifiers. Fractional values were calculated for mean agreement, with a value of 1 assigned to

	A	B	C	D
Kana’a	178	154	108	89.5
Wepo	193	19	58	169.5
Black Mesa	538.5	536.5	738	503
Sosi	643.5	527	804.5	630.5
Dogoszhi	221.5	255	292	260
Flagstaff	382	328	234	260
Wupatki	96	3	5	30
Tusayan	542	322.5	630.5	514.5
Kayenta	72	98	55	131
Other	81	8	34	0
Undifferentiated	128	819	115	467

Table S-I – Total Tusayan White Ware type counts for classifiers A-D, from the full set of 3064 sherds.

an exact match, and 0 for no match. For partial matches, where one individual selected a single type (e.g. Black Mesa), and the other a split type (e.g. Black Mesa/Sosi), a value of ½ was assigned. Results for all classifications are shown in Table S-II. The highest level of agreement between any two classifiers was 0.626, with the lowest level at 0.524. These values are well below those from the Fish studies (1976, 1978) of 0.70 – 0.78 for four different classifiers trained in the same environment with a dataset of 90 Tusayan White Ware and Gray Ware sherds. Using Fish's phrasing, these results are "somewhat disconcerting" and suggest that there has been some drift away from the Tusayan White Ware consensus that existed during his study more than 40 years ago.

1.3 Identifying A "Consensus" Dataset For Use in Training The CNN Models

The lack of agreement between pairs of human classifiers meant that any labeled dataset created from their classifications was likely to have a substantial amount of error built into it. To minimize this, it was decided to identify "consensus" sherds, ones for which the combined results from the classifiers allowed a reasonably confident type to be assigned. This would provide the Convolutional Neural Network models with a set of labeled data that it could be trained and tested on with at least some level of confidence. The following rules were employed to create a dataset of consensus sherds:

- Sherds were assigned "votes" using the numeric criteria given previously (value of 1 for a single type only, value of ½ for mixed types).
- A sherd that received a vote total of 2.5 or higher from the human classifiers (out of a maximum of 4) for a specific type was automatically assigned that type.
- Sherds with at least two votes, and a plurality over other types were assigned that type.
- Sherds with tie votes were eliminated from the consensus roster, with the exception for Wepo described below.
- Using the above criteria, only 53 sherds were classified as Wepo, a very small fraction of the total number. In part, this was likely due to classifiers B and C assigning the Wepo type to far fewer sherds than classifiers A and D, possibly reflecting some lack of confidence in this type; classifier B in particular expressed serious concerns about the vague criteria for identifying sherds as Wepo. This low number could negatively affect the ability of the CNN models to identify the design features associated with this type.. To compensate for this, and boost the total number of Wepo sherds, the criteria for Wepo were loosened slightly. Ties between Wepo and another type (always either Kana'a or Black Mesa) were broken in favor of Wepo. If the sherd had at least 1.5 votes with a plurality in favor of Wepo, the sherd was labeled as Wepo.
- Sherds that were typed by two or more individuals, but received no consensus, were classified as "Mixed", and not used for training the CNN model.
- Sherds that were only classified by one individual were classified as "Single" and not used for training the CNN model.
- Sherds identified by all 4 classifiers as "Indeterminate" were classified "Indeterminate" and not used for training the CNN model

Using these criteria, 2407 sherds were assigned a consensus Tusayan Whiteware type (Table S-III) out of the original 3064, or 78.6%. Of these, approximately 85% received 2.5 or more

Classifier	A	B	C	D
A	-	0.524	0.614	0.588
B	0.524	-	0.586	0.626
C	0.614	0.586	-	0.620
D	0.588	0.626	0.620	-

Table S-II – Agreement between pairs of human classifiers A-D on the original 3064-sherd dataset, expressed as a fractional value (0 = no agreement, 1 = complete agreement). Same as Table II in main paper.

Type	Consensus count
Kana’a	129
Wepo	89
Black Mesa	509
Sosi	604
Dogoszhi	261
Flagstaff	214
Tusayan	517
Kayenta	84
Mixed	556
Single	86
Undifferentiated	15

Table S-III- Consensus sherd counts by type votes, so the overwhelming majority of sherds were assigned a type with a reasonable level of confidence. The remaining sherds were labeled:

- Mixed, meaning two or more type labels were assigned, but no consensus as to type
- Single, meaning only one of the human classifiers felt comfortable in assigning a type
- Undifferentiated, meaning none of the human classifiers felt comfortable in assigning a type

Table S-IV shows the level of fractional agreement between the individual human classifiers and the consensus dataset types, ranging from 0.736 to 0.869. While substantially higher than the values for the original dataset (Table S-II), they are inconsistent with each other, and also lower than desirable. Additionally, the pairwise agreements between human classifiers on the consensus dataset (Table S-V), while also higher than the values in Table S-II, are lower than the values on the consensus dataset (Table S-IV). It is worth noting that although each human classifier labeled substantial numbers of sherds as “undifferentiated”, ranging from 115 to 819, only 15 sherds were labeled by all four human classifiers as “indeterminate/undifferentiated.

These results indicate that there is some level of uncertainty for type assignments “baked” into the consensus dataset, due to disagreements between the human classifiers on the proper classification of the dataset. This in turn is likely to limit the accuracy achievable by a CNN model trained on this dataset to a value comparable to that of the human classifiers. Any accuracy value achieved by the CNN model that is within the range of the human classifiers on the consensus (0.736 to 0.869) could reasonably be considered to be a promising result.

2. Selecting and Training The Convolutional Neural Network Models

2.1 Model selection and modification

The VGG16 (Simonyan et al. 2015) and ResNet50 (He et al. 2016) convolutional neural network models were chosen as the baseline for this work. Versions of both VGG16 and ResNet were past winners of the Imagenet Competition (Gershgorn 2017), achieving high scores for classifying test images into one of 1,000 different categories. One practical reason for choosing VGG16 and Resnet50 is that their simpler structures make them tractable for personal computers with reasonable performance specifications. Thus, while not currently considered state-of-the-art, their scores are only a few percentage points below state-of-the-art, provide acceptable results, and do not require extraordinarily powerful computational hardware. The models used in this study were processed on a Windows PC system with the following specifications:

- 6-core / 12-thread Intel CPU
- 64 GB of DDR4 RAM
- nVidia-based GTX1080 graphics card with 8 GB dedicated RAM

Of these, the most important hardware specification is the graphics card. Modern machine learning frameworks can accelerate performance dramatically by using nVidia-based graphics cards to speed up the numerical calculations required for neural network training. We would recommend at least an nVidia GTX980 or higher, with at least 8GB of RAM, as the minimum acceptable specifications. The Google Tensorflow machine learning framework optimized for

Classifier	**Consensus agreement**
A	0.797
B	0.736
C	0.869
D	0.833

Table S-IV - Agreement between human classifiers and baseline consensus 2407-sherd dataset, expressed as a fractional value (0 = no agreement, 1 = complete agreement). Same as Table III in main paper.

Classifier	A	B	C	D
A	-	0.604	0.709	0.680
B	0.604	-	0.668	0.649
C	0.709	0.668	-	0.727
D	0.680	0.649	0.727	-

Table S-V –Agreement between pairs of human classifiers on baseline consensus 2407-sherd dataset, expressed as a fractional value (0 = no agreement, 1 = complete agreement). Same as Table IV in main paper.

use with nVidia CUDA libraries was employed for the model training and evaluation; the Python-based Keras front-end used to simplify model creation and training.

Both chosen models have openly available pre-trained model weights from the 1,000-category Imagenet competition dataset. This allows for use of “transfer learning” (Yosinski et al. 2014), where pre-trained model weights from a classification problem (image recognition in this case) can be used to “jumpstart” the training process for a similar classification problem. Classification using multiple versions of the same or different CNN models allows combining results from both in an “ensemble” approach, which has been shown to improve accuracy on image classification problems (see Krizhevsky et al. 2012; Simonyan et al. 2014; He et al. 2016).

The model used for transfer learning (either VGG16 or ResNet50) was loaded into the computer program, using the weights for the pre-trained Imagenet model on which it was trained (1000 different image categories (Russakovsky et al, 2014)). Generally, each model can be thought of has having two sections. The initial stage, into which the image data is first loaded, contains the code used to extract information about features important in classifying images. The latter stage (“head”) contains the section that takes the feature information and uses it to classify the image. Since this latter section is specific to the set of images to be classified, in “transfer learning” it is removed and replaced with a section that matches the image types being trained; additional layers can also be added to optimize the model for the particular image recognition problem. For the sherd classification problem, based on some experimentation, the following layers were substituted for the standard head section of the two model types used:

VGG16

- Flatten
- 3 x the following structure
    - 512-element Fully Connected layer, RELU activation
    - Batch Normalization
    - 50% Dropout
- 8-element Fully Connected layer, softmax activation for final classification of sherd images into confidences for the 8 sherd classes

ResNet50

- GlobalAveragePooling2D
- Flatten
- 512-element fully connected layer, RELU activation
- 8-element Fully Connected layer, softmax activation for final classification of sherd images into confidences for the 8 sherd classes

Source code for training both types of models is available.

2.2 CNN Model Training Flow

The consensus dataset described in section 1.3 was used for training the CNN models. A standard approach for training and evaluating the accuracy of CNN models is to do a stratified split of the labeled data into training and test datasets. Stratified means that the proportions of labeled data types are the same between the training and test datasets. The models are trained on the training dataset alone, with the test dataset used as an independent monitor of accuracy. Using the full dataset for training will often result in a CNN model with high accuracies on the training set, but which may not necessarily obtain high accuracies on independent datasets. In addition, another standard approach for evaluating machine learning models, stratified cross-folds, was employed. Four of the folds were combined into a single dataset used for training the model, with the fifth fold used as the evaluation test set. This process was repeated an additional four times, each time using a different fold for the evaluation test set. This allowed independent evaluation of all sherds in the dataset, as well as ensuring that the accuracy results weren't skewed negatively or positively because of an unfavorable/favorable data split.

One potential issue with training CNN models is "overfitting", where the model adapts itself to produce high accuracies on the training data, but which may not necessarily produce high accuracies on test data. One method to address this problem is to maximize the amount of data available for training. While reasonable numbers of images were available for 5 of the sherd types (Black Mesa, Sosi, Dogoszhi, Flagstaff and Tusayan), there are substantially fewer images for Kana'a, Wepo and Kayenta. To address this, we used "data augmentation," where the images were randomly modified to increase the variability of images used in training. For training the sherd models, the image set was augmented using two techniques. First, during each training cycle ("epoch"), every image was rotated by a random angle between +/- 180 degrees. Since it is generally difficult to determine the "correct" orientation for sherd designs unless part of the vessel rim is present, it was anticipated that this would improve the overall ability of the model to classify sherd types regardless of the orientation of the original image. Second, images were enlarged or shrunk by a random zoom factor between -30% (reduced) and +30% (enlarged). For training the two CNN models (VGG16 and ResNet50), the sherd images needed to be resized to 224x224 pixels. For large sherds, downsizing could lead to diminution of relevant details; for small sherds, details could be exaggerated due to upsizing. Varying the zoom factor allows for detail recognition over a wider range of scales. Images were also converted from the original color to grayscale, to remove any effects from sherd discoloration from soiling or weathering, or from variations in white balance when photographing the sherds. Since both VGG16 and ResNet50 require full color RGB images, the grayscale sherd images were converted back to full color before being used for training.

The following process flow was used for generating trained CNN models:

1. For preliminary training, the weights of the initial stage are "frozen", while the weights of the head stage are set to random values. This prevents the weights of the initial stage from being modified during preliminary training, which can detrimentally degrade the utility of the previously-trained model weights of the lower section.
2. The model is then trained with all the training images run through 10 cycles (called "epochs") to refine the weights in the head stage. After this preliminary training is complete, the weights for the entire model are unfrozen, and the final training commences. Each model was trained for 120 "epochs", where each epoch represents the processing of a full set of images from the training set.
3. After each epoch, the accuracy of the model was evaluated using the test dataset. If the test accuracy exceeded the best previous results, the model was saved as a checkpoint. The final saved model was the one that had achieved the highest test accuracy. Figure S-2 shows a typical plot of training and test accuracy as a function of epoch number for

the ResNet50 model. While 120 epochs of model training were run, test accuracies typically stabilized after approximately 60 epochs. Note that the test accuracy in Figure S-2 remains flat over the last 60 epochs. If the CNN model was overfitting to the training data, this would manifest as a drop in the test accuracy; the fact that the test accuracy did not drop is usually a sign that overfitting was not a problem.

4. For each cross-fold, three training runs were done for each CNN model (VGG16 and ResNet50), and five sets of stratified cross-folds were evaluated.
5. The sherd images from the test dataset were evaluated with the 6 CNN models created using each train/test fold set (3 VGG16, 3 ResNet50) to generate type predictions, and the results from all 6 averaged for each cross-fold (ensemble evaluation). The results were output as a data file with the image name, the consensus type, confidences from the model that the image belonged to any of the 8 ceramic types, and the predicted type (the type with the highest confidence value). Results from all five cross-folds were collated into a single data file. A random sample of the CNN model output is shown in Table S-VI.

2.3 Metrics For Evaluating the Output of the CNN Models

The output data from the last section described above formed the basis for summary statistics reporting:

- **Top-1 Accuracy:** The overall level of agreement between the consensus and both the types predicted by the CNN model and the types predicted by the human classifiers
- **Fractional agreement:** Pairwise comparisons of the type classifications for both the CNN model and the human classifiers
- **Confusion matrix:** A matrix with predicted type as rows, and consensus types as columns. This shows how many predicted sherds matched the consensus type, how many did not match, and which types they were assigned.
- **Classification report:** A table of precision and recall for each of the sherd types. Precision is the ratio of sherds correctly assigned a particular type divided by the total number of sherds assigned that type (correctly and incorrectly). Recall is the total number of accurately typed sherds for a particular type, divided by the total number of sherds that actually had that consensus type.

The full process flow, including both CNN model training and evaluation, is shown schematically in Figure S-3.

Image file	Consensus	Predicted	Kana’a	Wepo	Black Mesa	Sosi	Dogoszhi	Flagstaff	Tusayan	Kayenta
23346_42.jpg	Kana’a	Kana’a	0.31	0.275	0.135	0.113	0.037	0.11	0.015	0.003
23360_02.jpg	Wepo	Kana’a	0.456	0.37	0.094	0.016	0.02	0.033	0.009	0.003
23345_18.jpg	Black Mesa	Black Mesa	0.002	0.005	0.95	0.04	0	0	0.002	0.001
23345_30.jpg	Black Mesa	Sosi	0.001	0.001	0.412	0.58	0.001	0.002	0.003	0.001
CD_Tusayan_322.jpg	Tusayan	Black Mesa	0.006	0.018	0.456	0.053	0.012	0.016	0.382	0.057

Table S-VI – Sample lines of datafile containing results from CNN model processing of test datasets.

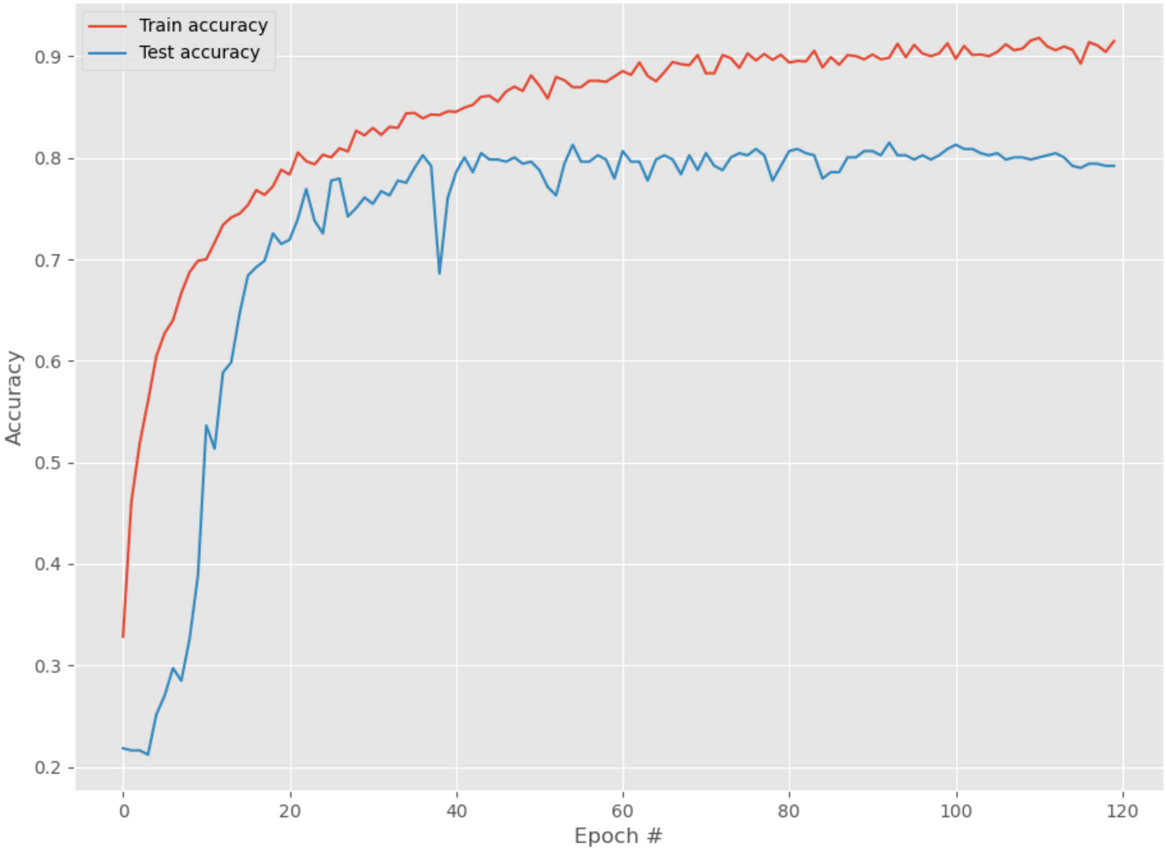


Figure S-2 - Typical plot of accuracy for both the test and train datasets as a function of epoch (cycle of image inputs), using the ResNet50-based model. Test accuracy typically flattened out at around 50-60 epochs. If the test accuracy had started dropping, that would have been a sign of “overfitting”, i.e. the model learning to classify the training dataset well at the expense of being able to generalize to the test dataset. No signs of overfitting were seen training any of the models.

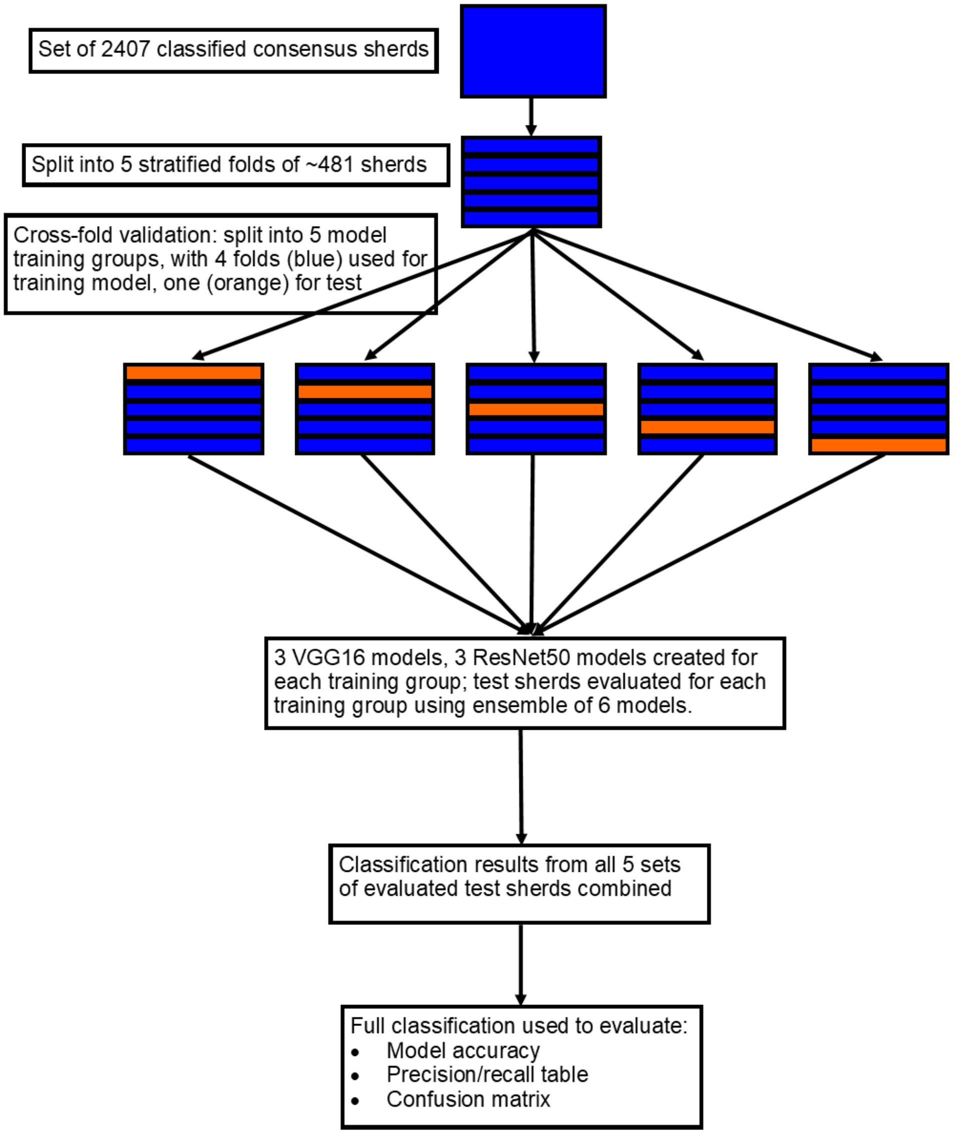

Set of 2407 classified consensus sherds
Split into 5 stratified folds of ~481 sherds
Cross-fold validation: split into 5 model training groups, with 4 folds (blue) used for training model, one (orange) for test
3 VGG16 models, 3 ResNet50 models created for each training group; test sherds evaluated for each training group using ensemble of 6 models.
Classification results from all 5 sets of evaluated test sherds combined
Full classification used to evaluate:
Model accuracy
Precision/recall table
Confusion matrix

Figure S-3 – Schematic flow diagram of CNN model data processing, training, and analysis. Set of 2407 consensus sherds is split into 5 stratified folds, each with the same fraction of Tusayan White Ware types. 5 sets of models are created, with 4 folds used for training the model, and one held out as a test set to monitor accuracy during training. The 6 CNN models created by each train/test set are used as an ensemble to classify the sherd types in the test set. All off these test set results are aggregated into the full dataset, where they are compared to the consensus types to evaluate accuracy, precision/recall, and the full confusion matrix for the CNN models.

References

Fish, P. R.

1976. Replication studies in ceramic classification. Pottery Southwest 3(4): 4-6.

1978. Consistency in archaeological measurement and classification: A pilot study. American Antiquity 3(1): 86-89.

Gershgorn, D., 2017. ImageNet: The Data That Transformed AI Research – And Possibly The World. Quartz Magazine (online), https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/ (accessed 1/2021).

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. NIPS proceedings, pp. 1106–1114.

Russakovsky, O., Deng, J., Su, H., Krause,J., Satheesh, S., Ma,S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A., Fei-Fei, L., 2014. ImageNet Large Scale Visual Recognition Challenge. https://arxiv.org/abs/1409.0575 (retrieved 1/2021).

Simonyan, K., Zimmerman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015. https://arxiv.org/abs/1409.1556.

Yosinski, J., Clune, J., Bengio, Y., Lipson, H., 2014. How transferable are features in deep neural networks? NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems – v. 2, December 2014, pp. 3320–3328.