Rock Art Classification Through Privacy-Guaranteed Ensemble Machine Learning

Ovidiu-Emilian Buțiu[1]

Artificial Intelligence and Distributed Computing
West University of Timișoara
Bulevardul Vasile Pârvan nr. 4, Timişoara Timiş România 300223
ovidiu.butiu01@e-uvt.ro

Abstract. The use of digital media provides important benefits in preserving, but also gaining an insight into cultural or natural heritage, such as rock art practices in the past and the present. With this digitized data, a constant need for assisting AI software in speeding up the monotonous parts of deciphering rock art is created. This paper reviews the current state and the existing problems in the field of rock arts, which is in the domain of digital heritage. The discovered main issues are the high variety and high volume of data, which goes unused due to various circumstances. We compare different approaches involving individual models like Faster R-CNN and ensemble frameworks containing variants of the architectures Resnet, Resnext, and Densenet with voting systems, like PATE. We experiment with the most promising find PATE, which offers a privacy-assured ensemble of models, each one of them trained using local data. The conclusion gained after the analysis of various papers and the implementation of the closest approach to fulfilling our needs, PATE, is under-performing with an average accuracy of around 10-15% We believe the main cause is the small and unbalanced dataset, which could be further artificially increased in future research through GANs and transformations.

Keywords: Rock Art · Classification · ResNeXt64 · Resnet18 · DenseNet201 · Voting Mechanism · Ensemble ML · PATE · Knowledge Distillation.

1 Introduction

Rock art is currently studied around the world, with almost every country providing research sites of the past. These sites can be of different context types, a few discovered examples being trophy rooms, and shelters used for initiation rituals and day-to-day activities [14]. They also vary in visual depiction due to cultural differences [5], outsider and environmental damage[14], or used tools, all of these mainly depending on the region. Many teams are looking into these shelters, trying to uncover their rich past, each with huge volumes of data to analyze and interpret.

Currently, the largest body of evidence of the beginnings of humanity's culture is represented by prehistoric rock art. It has provided a profound influence

when it comes to the beliefs and cultural conventions of consequent societies up to now. As such, it is an important part of the collective memory of humanity and also a most significant and enduring insight into our cultural evolution.

The problem we look to solve is the classification of these rock art motifs using ML algorithms. For possible solutions, we not only look into individual model approaches, but also ensemble ones and ones that provide dataset privacy based on identified challenges described in the problem formulation.

This paper aims to provide a better understanding of the apparent obstacles to the task of classifying rock art. We believe that the main challenges stem from data heterogeneity and imbalance in the proposed solutions we've gathered from other research papers. This is shown in different papers as the rock art dataset they use is smaller, usually under 200 images of motifs, and tends to feature an imbalanced, yet varied amount of simple and complex symbols.

Through our exploratory research, the gained insight could stand as the foundation of various future improvements for both image classification of heterogeneous datasets, and for the development of AI software assisting in rock art classification. With the use of the PATE framework, there is the apparent possibility of also establishing a distributed ensemble of various privately trained models, which could be used cooperatively by researchers in the competitive domain of rock arts, where a lack of trust among peers is common, caused by fear of having ideas, or even personal unpublished work stolen, as an already published paper is considered the only way to prove one's ownership in the fight against plagiarism. This leads us to form the following research questions:

- What models used in image classification tasks obtain a satisfying performance for rock art?
- What are the encountered difficulties when dealing with rock art motifs?
- How much does using an ensemble of ML models instead of individual models impact performance?
- Would a mix of different models be more beneficial to fit the student model in our approach, or would a homogeneous ensemble provide better results?

For the remainder of the paper, we look deeper into the problem we are focusing on, the state-of-the-art concluded from related works, followed by describing our contribution. Afterwards, we describe possible ethical concerns. We follow with the experiments taken to answer the offered questions, prove the efficiency of our solution, and provide a thorough comparison with the prior researched similar approaches to our problem. Finally, we offer a conclusion reiterating the problem, alongside a short summary of our solution and its achieved results. A short list of future directions which we plan to pursue next is also provided at the end.

2 Problem Formulation

The main problem of rock art classification is having to go through this large amount of digital data with the tasks of identifying and finding the context for

each apparent motif, which results in an overall very repetitive and tedious loop. The proposed solution for this classification problem is the use of ML algorithms. This approach saves a small amount of time for each identified motif, which could range from one or two to tens for every image. In turn, it would overall save a generous amount of time spent on identifying lots of data.

At the same time, the proposed solution must deal with the difficulties caused by data heterogeneity, as motifs can range from simple shapes to complex and varied depictions. This is noted by another similar approach [7], which experienced poor results when dealing with motifs depicting people, due to their heterogeneous nature.

This points us towards the use of a varied dataset alongside ensemble ML, which offers the possibility of using varied models together, each trained on a subset of data, which in our case would be the images split into countries of origin. This would also mean a better performance than simply using a generic ML algorithm. The problem with this approach is that it necessitates a privacy guarantee for the datasets used in training and testing, due to either the lack of

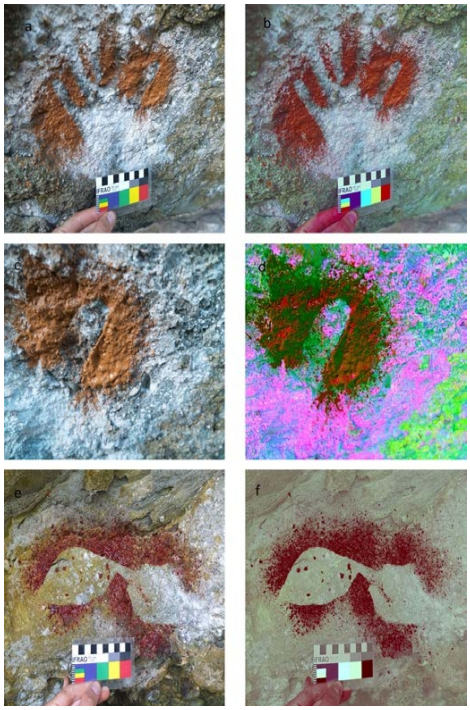


Fig. 1. Examples of various stencil types discovered at Apuranga site in 2018. *"Left down (a, c and e) before enhancement: a hand/palm, a thumb, and unknown leaf stencils, respectively, with the appearance of being recently made. Right down (b, d, and f) after picture enhancement using DStretch (yre, crgb, yre color filters, respectively) shows no older art (original photos: William Pleiber, Papuan Past Project)."* [14]

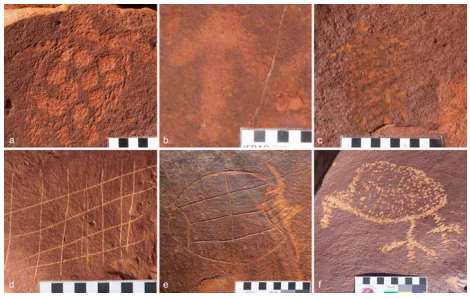


Fig. 2. Examples of various motifs, some harder to spot the shape of, created using different techniques: *"(a) pecked turtle showing two differently sized impact shapes ; (b) pounded bird track motif, probably produced with a rounded hammerstone; (c) abraded turtle motif; (d) scratched grid; (e) incised and abraded turtle; (f) pecked and gouged bird."* [5]

connections and trust between researchers, or legal or ethical constraints, which makes procuring and using data difficult.

To provide an ensemble ML approach with a privacy guarantee, which would allow the usage of private data in training and testing, we look towards Private Aggregation of Teacher Ensembles (PATE) [12]. This framework transfers knowledge from an ensemble of teacher models, trained on partitions of the data, to a student model. This protects the privacy of the training data used during learning and also offers the freedom of picking any suitable learning technique for data to use for the teachers.

At the same time, partitioning the data avoids overlapping, which produces a variety of models that independently predict labels. PATE also offers easy scalability, allowing various devices, each with its models trained on private data as stated previously, to work together as a bigger, better-performing model. This feature can offer cooperation between various research teams, which can now not only keep their data private but also assist each other by adding a model trained on their dataset, possibly improving performance and preventing biases from forming as more varied data is added overall.

Our goal is to achieve a satisfying performance in motif detection and classification, which is why we look into ensemble ML, specifically the PATE framework's potential to be used with various models to improve performance in identifying simple and complex motifs when dealing with rock art images from various countries. For that, we must also look into how well models perform when it comes to rock art classification. As PATE allows us to fit a student model using sensitive data, there is a need to apply noise while training, which may heavily impact performance, which is why a baseline would be useful using the teacher model performance. This would also provide insight for improvements and allow comparison of the performance of the trained student model.

3 Theoretical Background

According to Robert Bednarick [3], rock art is defined as human-made markings on natural rock surfaces with the use of additives (pictograms by applying material) or through a reductive process (also defined as petroglyphs, created through the removal of rock surface). The former regards rock paintings, pigment drawings, stencils, and beeswax figures, while the latter term may cover engravings, percussion petroglyphs, and finger flutings. These markings occur in nearly all countries, although in an uneven distribution, which can be attributed to cultural convention differences and a result of preservation bias (taphonomic attribute). An example of rock art can be seen in Figure 1, which showcases stencils discovered at Apruanga site in the year 2018, depicting a hand, a thumb and respectively leaves. Another example can be seen in Figure 2, which instead depicts different techniques, including carving, used to display objects and animals.

4 Methodology

The papers selected in this report have been mainly selected using research platforms such as Arxiv and ResearchGate, or through the searching tool Google Scholar.

The keywords used for searching for these papers are *ensemble ml*, *federated learning*, *rock art classification*, *rock art taxonomy*, *heterogeneous*, *PATE*, *rock art ml*, *privacy ml*, *individualized privacy ml*, *digital heritage ml*, *ml security*. The focus was on gathering papers on various ensemble ML models to be compared when used in the taxonomy of rock art, but this required more information on said domain. Thus, we also gathered some keywords to look into more information about the domain of digital heritage and more specifically the classification of rock art. We prioritized the context when it came to filtering the found ML approaches, as we wanted to look into what has already been tried for the task of rock art classification. If that was not the case, like for the ensemble ML approaches, we simply focused on the most recent and high-impact work to be used for this topic.

The datasets we decided to use for these experiments consist of images from various public sites in the countries US, Spain, Italy, Bulgaria, and Armenia. These count up to a total of 257 images. We have cropped all the labels from these images, resulting in 948 separate images of various labels. The data, as seen in Table 1, is both small and poorly balanced. This is due to the lack of sources and some of the motifs being more commonly featured than others, whether to tell a story or to paint a picture. There is the possibility of the dataset forming a bias towards the labels “Person", “Goat" and “Circle" as they form the majority of the motifs. This imbalance could lead the model to mistake stags for goats, for example, as they are similarly drawn, with differences in the representation of their horns.

The same mistake could happen when it comes to the detection of spirals, as some of these motifs are represented inside circles, which the model could be

Table 1. Statistics of the private rock art used data

Country	*Count*	*Train Split*	*Test Split*	*Person*	*Goat*	*Stag*	*Circle*	*Spiral*	*Zigzag*	*Cross*
Armenia	106	74	32	91	132	18	53	9	27	10
Italy	47	33	14	83	11	50	39	3	3	4
US	91	64	27	108	44	4	104	58	66	31
Total	244	171	73	282	187	72	196	70	96	45

biased towards. We believe that this imbalance could be solved or at the very least improved through the use of Generative Adversarial Networks (GANs), as we could use the already procured data as input for new artificially generated data.

We believe the representations of the “Person" label to be complex enough not to cause bias, but even so there may be the slight possibility where it could mistake some of the animal representations as human, as they may have similarities in patterns.

One such example would be the horns being mistaken as the spears held by some of the painted humans. It could even happen to simpler models, like spirals or circles, which could be mistaken as the heads of these human figures.

For data pre-processing, we previously used an application named labelImg to provide an XML file with the labels for each image for an earlier unsuccessful attempt at the task of object detection. We used these XML files to extract the cropped images, featuring only one motif each instead for our image classification task.

We have also applied other transformations to the images besides just resizing in an attempt to increase the dataset's size artificially. We applied 60-degree rotations six times, each time adding the transformed data on top of the dataset. We've also attempted to change the hue, saturation, and brightness to better differentiate the rock carving/painting from the stone itself.

Rock art is considered sensitive, which is one reason for the lack of data. It is also highly important to some archaeological research. One solution to this is the use of approaches that provide a privacy guarantee, such as Private Aggregation of Teacher Ensembles (PATE).

5 State of the art

5.1 Rock Art

Most of the found papers about the use of machine learning methods that look into the topic of rock art focused on image classification, similar to our problem. At the same time, those focused on ensemble ML used datasets that can be considered highly different in characteristics from the rock art dataset we used in our experiment.

In our work, we focus on rock art pictures and carvings representing various shapes, similar to those featured in Figure 1. These are also varied due to the

different countries of origin and their multiple regions as shown in Table 1, making the classification of some motifs a challenge due to the heterogeneity of the formed dataset.

Some of these works focus on the classification of various medical maladies, such as skin lesions, breast tumors, and brain tumors. The data used in these approaches contain similar challenges as the ones mentioned in the previous sections. These challenges are their unpredictable and complex nature, which is the most apparent issue when it comes to our datasets.

Some of these works focus on detecting whether the image contains rock art or not, or on classification of rock art from only a single country or region. These approaches tend to provide solid results for the test images featuring rock art found in the country of origin, but their focus is not on providing a more general model, and thus lack tests on data from other regions or countries.

The remaining works focus on

Even so, we have found this previously mentioned complexity issue on a smaller scale in a work using Faster R-CNN for 3D rock art data, for example, which also had a hard time classifying the complex labels with varied similar representations, such as “animal" or “human". We expect this to also be the case with our merged datasets, as the different representations would greatly affect the more complex labels, such as “deer", or “human".

Thus, the presented work only offers some likely useful approaches when it comes to ensemble ML, but does not provide enough information for our use case. We expect to achieve worse results than the ones described in this subsection.

We begin with a research paper delving into the identification and classification using Faster R-CNN of 3D rock art data processed into images [7]. Although this ensemble ML approach achieved a mAP of 32.47%, its results on the simple common labels such as boats were promising, achieving over 60% in precision for said class. This can be considered a minimal score we aim to achieve with the models we use for classification.

The rest of the objects suffered from under-representation and thus were harder to conclude results from. An important noticed issue to be aware of in our experiment was the inability to capture all objects in the data with their predictions, which may stem from the need to adjust the IoU threshold. This was left for future research as it was outside of the paper’s scope and required further evaluation by setting the value lower than the one used in the experiment, which was 0.7, looking into the possibility of a relation between lower threshold values and higher recall values.

Visual inspection of the results also showed clear difficulty when it came to the model predicting larger objects with one single bounding box. As they averaged overlapping bounding boxes of the same class, the blame may stem from the bounding boxes not overlapping to a sufficient degree and causing misalignment with the motif during the box’s plotting.

Also, the worst overall performing object class seemed to be “animal", mainly because of the considerable variation, possibly combined with the limited number of training examples. This could also be the case for other classes such as

“human", that suffered from less performance compared to object classes representing much simpler figures, like “boat", or “circle", since in rock art research, humans or other anthropomorphic figures tend to come in combinations of bodily features and associated objects, causing them to be complex and varied. We should be aware of this possible issue, as our labels feature human and animal figures.

We follow with another recent study on Australian rock art which use the models VGG, Inception, and ResNet in their approach [8]. These models achieve very promising accuracy and F1 scores of around 80-90%. Although this research does not focus on the classification of motifs, more explicitly just rock art detection, it proves that the challenge lies solely in identifying the different represented shapes.

These papers showcase overall promising results achieved in classifying rock art for a country, whether only for specific types of motifs or just simply for detecting possible motifs in an image, but a lack of focus on the bigger apparent issues when it comes to dealing with rock art. These issues are cultural diversity and lack of data privacy, as stated previously.

5.2 Ensemble ML

Next, we looked into what approaches deal with the problem of data heterogeneity, as we suffer the same challenge from the cultural diversity of rock art. This challenge stems from the fact that motif depiction may vary depending on region, which can cause data overfitting and a bias towards specific depictions of certain cultures. At the same time, the lack of data privacy is a big issue for researchers, as the competitive environment of rock art research means no trust in sharing datasets that may be used in the development of assisting tools for classification.

In our search, there was a lack of ensemble approaches that fit our topic. As a result, we looked into the topic of medicine, which uses datasets that share similar challenges to the ones we discuss about rock art datasets. One such trait is data heterogeneity, as some diseases can be represented by varied shapes that can be either simple or complex. Another trait is the need to protect the privacy of the dataset, as datasets from this topic may contain patient info that needs to be kept confidential. This is similar to the requirement to keep rock art datasets private, to allow the use of private datasets in training and testing.

Another found paper attempts to deal with skin lesion classification, which has similar traits to the data we are working with, as they are heterogeneous, varying in size, color, shape, and complexity [13], through an ensemble learning approach which combines three deep convolutional neural network (DCNN) architectures known as Inception V3, Inception ResNet V2 and DenseNet201. This is done to improve overall performance compared to a single DL model approach, producing promising classification performance, resulting in 97.23% accuracy, 90.12% sensitivity, 97.73% specificity, 82.01% precision, and 85.01% F1-Score.

Assiri et al. [2] proposed an algorithm for breast tumor classification using a voting mechanism, first selecting the 3 best-performing models out of 8 evaluated classifiers, based on the F3 score due to the importance of false negatives for this subject. The following winners, simple logistic regression learning, support vector machine learning with stochastic gradient descent optimization, and multilayer perceptron network, are then used for ensemble classification using various voting mechanisms. This results in the majority-based voting mechanism achieving an accuracy of 99.42% on the publicly available Wisconsin Breast Cancer Dataset (WBCD).

Kang et al. [10] focused on a different approach for the classification of brain tumors, which once again could share similar traits to our data due to their unpredictable and complex nature, by adopting the concept of transfer learning. CNNs are used to extract deep features from MRIs, which are then evaluated by ML classifiers. The top three best performing are selected, concatenated then fed once again into ML classifiers as an ensemble of deep features to predict the final outcome. After testing on three different datasets, this method proves promising for overcoming limitations of a single CNN model, the best performance being achieved by support vector machine (SVM) with radial basis function (RBF) kernel, especially on larger datasets, with the highest accuracy reached on each dataset being between 93-98%.

This related work provides some solutions that could work if implemented for the domain of rock art, but even so, there is still one previously mentioned issue that we have yet to go through, which is the inability to use sensitive data. The need for privacy guarantees also can happen in the domain of medicine, as patient data privacy limits the available data to be used for training and testing models.[12] Also, due to mainly focusing on only a specific type of affliction, these works did not have to account for high data heterogeneity, similar to the rock art research papers only focusing on one country.

Compared to these related works, we plan on using the PATE framework. This offers the possibility of training teacher models that can be on different computers, each with its private dataset, to work together through a voting mechanism and with applied noise to train a student model. This results in the protection of the data used by the teachers to train the student.

This approach's design is also adaptable and offers scalability, allowing any model to be added to the network as long as they are properly configured, meaning there is the possibility of compensating for the weaknesses of certain models through this ensemble. PATE offers a voting system that evaluates each teacher model's response and trains the student model using selected queries for which an overwhelming consensus was reached, passing over a threshold T.

With this method, we aim for promising results in improving rock art classification accuracy by offering the possibility of using or adding sensitive data to the mix to combat the lack of public data. The framework presents the possibility of using sensitive data and features of scalability and adaptability, meaning a path to cooperation between rock art researchers and ML can be paved.

If the resulting student model performs well, the PATE framework also offers us the possibility of further research into the side of data privacy, with the possible implementation of individualized privacy [4]. This approach may further improve the student model's performance, as the teacher models would be able to pick from a range of noise levels, depending on the required strictness for the privacy budget of the used dataset. There are two existent mechanisms for this approach, the addition of either achieving the aforementioned task of providing individualized privacy.

Being allowed to use various levels of noise, instead of the highest required by the dataset, would in theory mean a better performance. One problem is that they require an entirely different way to evaluate from the standard PATE, which is already challenging enough due to the lack of available implementations to use for the framework's appraisal.

The reason as to why an entirely different method of evaluation is necessary is caused by it has to be done for particular data points or groups of data points separately instead of the entire dataset. Even so, they may be worth pursuing for this possible performance improvement, as their individualized variant of the framework showed better results when tested against the original implementation of PATE [4].

For our experiment, using the standard framework is satisfactory, as we simply aim to showcase the promise of using the ensemble approach to achieve solid results when it comes to rock art classification. Pursuing this approach of individualized privacy is better left for future works, or if proven to be required for the domain of rock art, due to necessary high differences in the levels of privacy between datasets.

Another alternative to the PATE framework is to use knowledge distillation from a cumbersome teacher model or the ensemble of teacher models, in our case, to guide the training of a smaller student model with a different architecture. It is known that this smaller student model can achieve similar accuracy to larger teacher models with less computational expense [11]. This alternative is close to PATE's approach of training a student model using teacher models, but instead of using a voting system, we follow the key idea explained in one of the selected papers on this topic that attempts to mix PATE with this approach [11].

This proposal suggests knowledge is not only stored in the model's parameters but also in the probability vectors (soft labels) it produces. These soft labels provide additional information concerning the relative distinction between classes. The process requires training a large deep neural network (DNN) with a softmax output layer to allow the generation of class probabilities, which are then used in training the smaller DNN.

The mentioned paper seeks to adapt the PATE framework to use knowledge distillation to more efficiently guide student training, which would result in mixed training. This is to aid in the compression of the deep learning model, allowing on-device deep learning for mobile devices, and to provide training data privacy preservation. Its complex approach is out of scope for our use case, but the key proposals regarding knowledge distillation and its similarity to the PATE

framework suggest we could compare it as a solid alternative to PATE, as it may provide similar, or even better results. This is due to its much smaller computational expense and better performance when dealing with smaller datasets than PATE. Similarly to PATE, we aim to provide a privacy guarantee through noise application on the logits supplied by the teacher models.

To achieve differentially private knowledge distillation, we follow a similar approach to another paper that uses it in training a compact and fast neural network [15]. The paper proposes a new mechanism that applies noise to the knowledge distillation in batches, in a much more complex but efficient way. The batch loss provided by the teacher is clipped using the adaptive norm bound, then perturbed carefully for privacy preservation. We aim to use a similar, simpler mechanism that goes sample-by-sample to perturb the provided knowledge distillation to preserve their confidentiality. This is done for each teacher that is part of the ensemble.

Another approach would be circumventing the data scarcity and privacy issues through synthetic data. This data originates from computer-based generation and can be used the same as normal data for solving specific tasks.

Various modern synthetic data generators can be used for this approach, such as Generative Adversarial Networks (GANs) and Variational Auto-encoders which are deep learning structures, to agent-based econometric models [6]. Their recent exponential growth resulted in the development of open-source or commercial tools that can automatically generate high-quality and clinically realistic synthetic datasets [6].

As this type of data can also incorporate real-world data, this can be used as a proxy for its real counterpart to provide privacy guarantees while allowing the researchers to conduct various experiments or analyses [6]. This has been proven to work in healthcare, where synthetic data was used as a proxy in analysis for real-world data for large-scale health surveys to protect patient confidentiality [6].

As stated previously, classification tasks in healthcare use data sharing similar characteristics to rock art data. This means we could provide the privacy guarantee through the generation of synthetic data to be used in training the models and thus circumvent the need for privacy guarantee in our approaches through noise generation. This would mean any ensemble approach could be used for our task, as the synthetic data itself provides a privacy guarantee for the original data.

Unfortunately, this is not an efficient approach, as dealing with the issue of data scarcity is more important than replacing the already small dataset we procured. Instead, for our experiment, we focus on showcasing the effects of adding synthetic data on top of the existing data we are using. This data was generated by a GAN, using the dataset we split for each of the three teacher models. This addition could improve overall model performance for both PATE and knowledge distillation and aid in fixing the data imbalance we are currently facing with the used dataset.

Even so, synthetic data carries the risk of bias and low interpretability. Generated images may reflect class imbalance, perpetuating initial bias [6]. This may instead cause worse overall model performance and lead the AI system to be “blind" to data beyond their training sets, as they are unable to make impartial decisions and accurately represent the unrepresented classes. It offers the risk of overgeneralization and has the potential to create correlations that are non-existent or incorrect. These are dangers we must be aware of and attempt to avoid through transparent documentation that may assist in identifying actual and potential errors.

For the task of rock art classification, we ended up selecting and implementing two approaches, the standard PATE framework and knowledge distillation, for which we aim to use the same models and datasets. We also experiment with the addition of artificially generated data to aid in combating the data scarcity and improving model performance.

6 Experiment

6.1 Research

Before our experiment began, we studied an example providing a PATE implementation in PyTorch[9] based on a research paper studying it as a generally applicable approach to providing strong privacy guarantees for training data [12]. The implementation showcases the use of this framework for the problem of image classification on the MNIST dataset, which would mean that we have to swap the model they use with the ones we plan to study and also implement a custom class to be able to work on our datasets.

Thus, we concluded that we are limited to using PyTorch, as the best approach would be to use the same packages used by Syft, the library providing a black box implementation through which we can analyze the performance of PATE in our experiments. This would ensure that as long as the models are functional and the experiment results are promising, testing these models using PATE would consist of simply replacing the models and datasets in one of its various PyTorch implementations.

The specific models that were planned to be used in this experiment were ResNet18, ResNeXt64, and DenseNet201, as these were the closest available choices provided by our used technologies to the similar work using ensemble ML in the task of classification.

6.2 Setup

The code used in our experiment is made available on GitHub.[1] As a start, we aimed to provide a performance comparison of each of the selected models. We used Anaconda to set up an environment running Python 3.7, which was required to be able to install and import the packages that would be used in

[1] https://github.com/ovybe/paterockartsota/

this research. Most of the setup consisted of installing an older version of the Syft library, downloadable thanks to an example of code with PATE as its focus subject [1], due to the latest version having been remade and not containing a PATE implementation at the time this experiment happened.

Overall, our experiment used the technologies: PyTorch 1.1.0, TorchVision 0.3.0, cudatoolkit 9.0, protobuf 3.20.1, matplotlib, and Pillow (< 7). The used implementations are run on a setup consisting of a desktop featuring an RTX 3060 ti equipped with 4,864 CUDA cores and 152 Tensor cores, an Intel® Core™ i9-11900F Processor, and 32GB of RAM DDR4 with a speed of 3200 MHz.

We separately trained the selected models, at least 5 runs for each country dataset on a newer version of PyTorch, which offered superior training times, then saved them as older variants so we could load them into the older version of PyTorch we're using for the implementation of the PATE experiment. This would allow us to select the best-performing model out of the saved runs for each dataset to be used for training the student model.

After selecting the best-performing models out of the total 15 instance models (5 runs for each of the three possible models) on the datasets, we aimed to test the ensemble performance, through first doing an ensemble formed of the same models, with each of the chosen models, using their best-performing run from the datasets.

We follow up on this approach with a final run training the student model through a teacher ensemble formed out of the best performing model for each dataset. For the student model, we wished to compare the achieved results by training the same CNN architecture on all the various ensembles. From the selected models, we decided to use ResNet18 as it was the fastest to train, only taking 3 hours and 40 minutes on average for 100 epochs for PATE and x for Knowledge Distillation.

For the PATE and Knowledge Distillation implementations, we reused the US dataset as the student model's training and validation dataset. To keep data confidentiality for PATE, we applied noise to the dataset through Laplace Distribution. For Knowledge Distillation, we had to instead apply noise to the logits provided by the teacher models for the student model's training.

For PATE, we used the trained teacher models to go through the dataset and vote on the predicted class. The possible predicted classes featured in the dataset were Person, Goat, Stag, Circle, Spiral, Zigzag, and Cross. This would provide the dataset used by the student model, validated with the help of the teacher models. All that is left for the experiment is to train the student model using the aforementioned dataset and provide its evaluation results.

The approach in training is slightly different for Knowledge Distillation, where the student is simply aided by the provided teacher logits in its training. Here, the student is also trained using the US dataset and its evaluation results provided.

For data pre-processing, we resized the images to 180 by 180, mainly to provide faster training and no size differences between them, and also used nor-

Table 2. A summary of performance for the best performing individual teacher models trained for 100 epochs to be used for the ensembles.

Model	Dataset	Val Acc	Val Loss	Recall Avg	F1	Tr Time
ResNet18	US	≈ 0.76	≈ 1.1	≈ 0.56	≈ 0.56	1 hr
ResNet18	Armenia	≈ 0.8	≈ 0.76	≈ 0.75	≈ 0.65	1 hr
ResNet18	Italy	≈ 0.72	≈ 1.4	≈ 0.6	≈ 0.76	1 hr
ResNeXt64	US	≈ 0.78	≈ 0.85	≈ 0.6	≈ 0.63	6.385 hr
ResNeXt64	Armenia	≈ 0.82	≈ 0.7	≈ 0.76	≈ 0.76	6.038 hr
ResNeXt64	Italy	≈ 0.79	≈ 0.95	≈ 0.65	≈ 0.71	6.33 hr
DenseNet201	US	≈ 0.84	≈ 0.73	≈ 0.66	≈ 0.67	3.12 hr
DenseNet201	Armenia	≈ 0.85	≈ 0.57	≈ 0.83	≈ 0.83	3.131 hr
DenseNet201	Italy	≈ 0.76	≈ 1.0	≈ 0.68	≈ 0.69	3.116 hr

malize with a sequence of means: 0.485, 0.456, 0.406 and standard deviations: 0.229, 0.224, 0.225.

We used a mean of 0, a scale of 0.1, and ϵ of 0.2 to apply the Laplace noise distribution. We used Stochastic Gradient Descent for optimization with a learning rate of $0.001 * 2.82 \approx 0.00282$ and momentum of 0.9. For the loss function, we used cross-entropy. The batch size used for training was 4 and the student model was trained for 100 epochs.

6.3 Achieved Results

The training was done using PyTorch version 2.5.1, with similar data preprocessing applied to the procured country datasets. We trained an instance of each model on three built datasets, respectively US, Armenia, and one combined from the data of the remaining countries. As we trained multiple model instances, we selected the best performing ones to represent each teacher ensemble.

We offer a summarization of individual model performance at Table 2 and student model performance at Table 3. We also provide summarized comparison graphs for the student model evaluations for both implementations. For PATE, these can be seen in Figure 9 for ResNet18, 7 for ResNeXt64 and Figure 8 for DenseNet201. For Knowledge Distillation, Figures 4, 5, 6 provides the validation accuracy, training loss and respectively validation loss result comparison of all three model instances together.

The performance summary of these models can be reviewed at Table 2. For the ResNet18 homogeneous teacher ensemble individual models, we have an average accuracy of $\approx 0.7 - 0.8$ with an average loss between 0.6 and 1.2. The F1 score average was between 0.5 and 0.8, showing some struggle with detection of certain classes. These models took the least amount of time to train, averaging at around 1 hour for 100 epochs for each teacher.

For the model trained on the US dataset (Teacher 1), the evaluation statistics suggest the model had a hard time identifying crosses. It also slightly struggled

at spirals and stags where the F1 score marked around $0.6-0.7$, compared to the average outcome being over 0.8. We believe the reason for this model's poorer f1 score on this dataset is due to the lack of Zigzag rock art for testing data, which caused a score of 0 during its evaluation, affecting the overall average.

For the model trained on the Armenia dataset (Teacher 2), the evaluation statistics suggest the model had a hard time identifying stags. We believe this is caused by dataset imbalance, as the data features only 132 goat motifs, compared to the 18 stag motifs available. This possibly resulted in the model forming a bias for the goat motif, which features similar characteristics to the stag rock art. The rest of the classes were on average resulting in around $0.7-0.9$ for F1 score, which we believe to be promising.

For the model trained on the Italy dataset (Teacher 3), which was also featured the least data of the three datasets, the evaluation statistics suggest the model had a hard time identifying crosses, stags and zigzags. Once again, we believe one reason for this to be the lackluster data, featuring only 4 stag motifs, while at the same time featuring 44 goat motifs.

This could again result in the model building a bias for goat motifs and having difficulties differentiating between the two and thus the F1 score for the stag class averaging 0.5. We also believe that the cross and zigzag motifs featured by the procured data from public rock art sites in Italy may feature similar characteristics, similar to the goat and stag motifs, resulting in the poor average F1 score of around $0.4-0.5$ when it comes to these classes. The rest of the classes were once again on average resulting in around $0.6-0.9$ for F1 score.

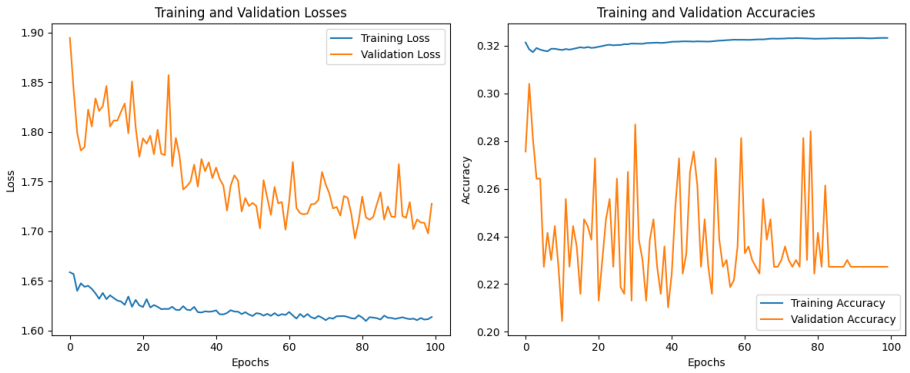


Fig. 3. A plot showcasing the training and validation losses and accuracy for the student model trained by the ResNeXt64 Ensemble using PATE on the US dataset.

For the ResNeXt64 homogeneous teacher ensemble individual models, we have an average accuracy of $\approx 0.78-0.82$ with an average loss between 0.7 and 0.95. The F1 score average was between 0.63 and 0.76, showing some struggle with detection of certain classes.

Most classes offered an average F1 score of over 0.8, with the models showcasing struggles again for crosses, stags and zigzags and an increased training time average for each model of 3 hours for 100 epochs.

Interestingly, the overall worst performing amongst the datasets seems to once again be the cross with ≈ 0.5 F1 score for US, ≈ 0.6 F1 score for Armenia and the lowest average ≈ 0.3 for Italy. The second worst is once again the stag motif, scoring an average ≈ 0.7 F1 score for US and Italy, but struggling at an average ≈ 0.6 F1 score for Armenia.

The individual models trained for the DenseNet201 homogeneous teacher ensemble offered the best results, with an average accuracy of $0.76 - 0.85$ and an average loss of $0.57 - 1$, but also took the longest to train, at an average of 6 hours for 100 epochs for each model. The average F1 score achieved by these models are around $0.67 - 0.83$.

Once again, the average model performance for each class seems to average around 0.8 or above, with the worst performance coming from the labels cross, stag and zigzag. For the US and Italy datasets, when it comes to the cross motifs, the trained models provide an average F1 score of ≈ 0.46. Once again, the stag features the second worst performance with an average F1 score of $0.5 - 0.6$ for Italy and 0.65 for Armenia.

The overall results from the individual models suggest that overfitting is occurring, as the training results showed great performance, with the validation results showing promising but much lower performance with lots of fluctuation during the model evaluation. We also noticed the models' struggle caused by the dataset imbalance when it comes to the stag, cross and zigzag motifs. We believe the cause, as mentioned previously, to be due to the model having a hard time differentiating these labels properly.

A possible solution to this issue would be procuring more data, which is difficult when it comes to rock art as mentioned before. Another solution would be using GANs to generate more labels. This could provide the model more data with similar characteristics to allow better differentiation between classes. Finally, we could focus on tweaking the models further and experimenting with data pre-processing, finding ways we could allow for overall better differentiation for motifs.

When it comes to the homogeneous ensembles, we trained a student model instance for 100 epochs for each set of teachers used on both the PATE and Knowledge Distillation implementations. We used an ε of 3.0, so the implementations would offer some privacy guarantee but also not be affected too much by the noisy data.

For Knowledge Distillation, the ResNet18 ensemble trained student model performed the worst, reaching a validation accuracy average of 0.2569 and an average validation loss of 2.2012. The ResNeXt64 ensemble trained student model reached a better validation accuracy average of 0.2847 and the best scoring average validation loss of 1.8195. The DenseNet201 ensemble trained student model offered the best accuracy, with an average of 0.3056 and an average validation loss of 2.1468. For a better overall view of the achieved results, we recommend

checking the graphs comparing their validation accuracy in Figure 4, their training loss in Figure 5 and their validation loss in Figure 6.

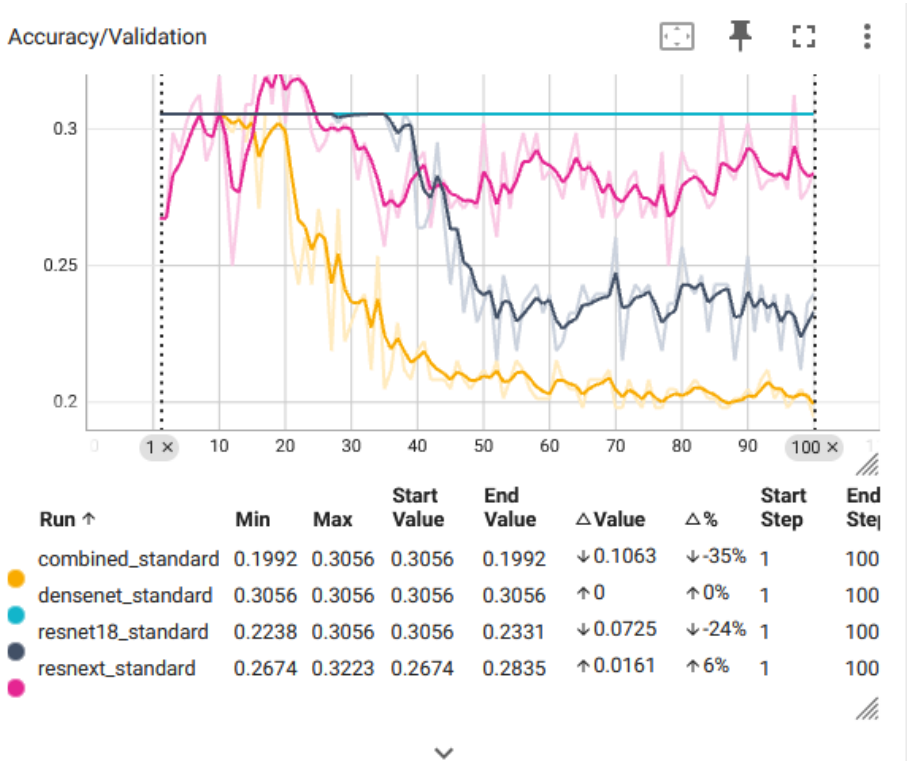


Fig. 4. A plot showcasing the validation accuracy for the student model trained by the ResNet18, DenseNet201, ResNeXt64, Combined (DenseNet201 for US, DenseNet201 for Armenia and ResNeXt64 for Italy) ensembles using Knowledge Distillation on the US dataset.

For PATE, the ResNet18 ensemble once again offered the worst performance, as can be seen in Figure 9 with a validation accuracy average of 0.2367 and an average validation loss of 1.7080. ResNeXt64 offers minimal improvements in validation accuracy performance, as can be seen from Figure 7 with an average of 0.2384 and an average validation loss of 1.7551. DenseNet201 once again offers the best performance as deduced from 8 with a validation accuracy average of 0.3183 and a validation loss average 1.6513.

We also trained the student model for each implementation using the best performing models from each dataset, with an ensemble formed of the best performing teacher models for each dataset, DenseNet201 for US and Armenia, and ResNeXt64 for Italy.

For Knowledge Distillation, this resulted in the model only performing slightly better than the ResNet18 homogeneous ensemble with a validation accuracy average of 0.22569, and slightly better than the DenseNet201 homogeneous ensemble with the average validation loss being 2.00693. For PATE, the student model's performance was close to on par with the DenseNet201 homogeneous ensemble, which was the best-performing one of the group, with a resulted average validation accuracy of 0.30159. The average validation loss was also the lowest out of all the ensembles, with a score of 1.63494.

Overall, we see PATE offering a better performance compared to Knowledge Distillation for the same teacher ensemble. Even so, the accuracy of the models is unfortunately lacking, as we believe the cause to be the student model suffering from underfitting.

This is due to the training and validation accuracy and loss being close in value, with fluctuations for the validation results. More training could provide better results, but the most significant improvements could come from an increase to dataset size. The lackluster datasets we procured are one of the main

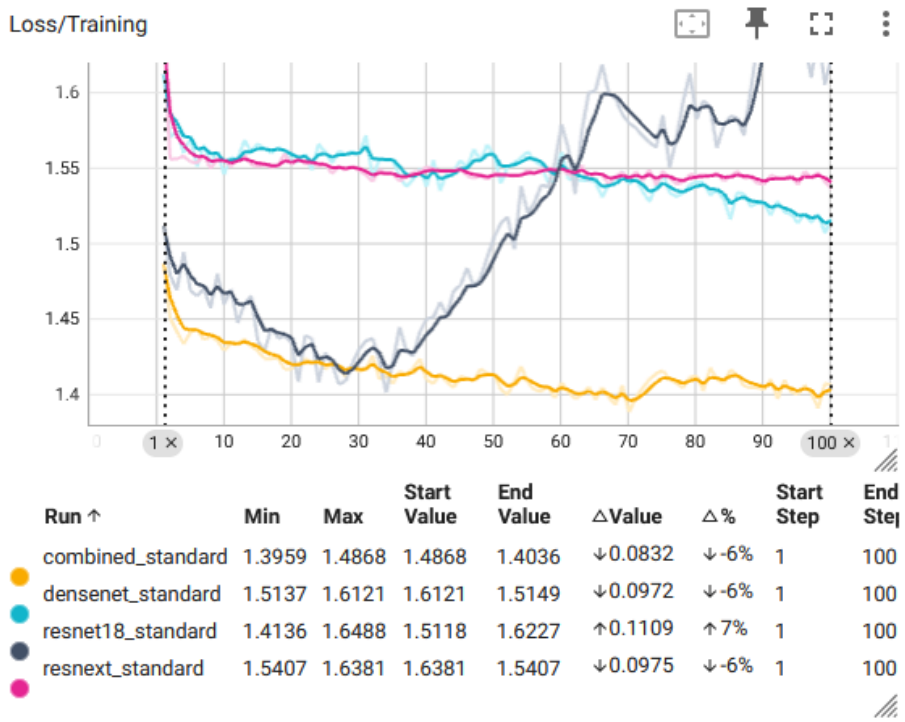


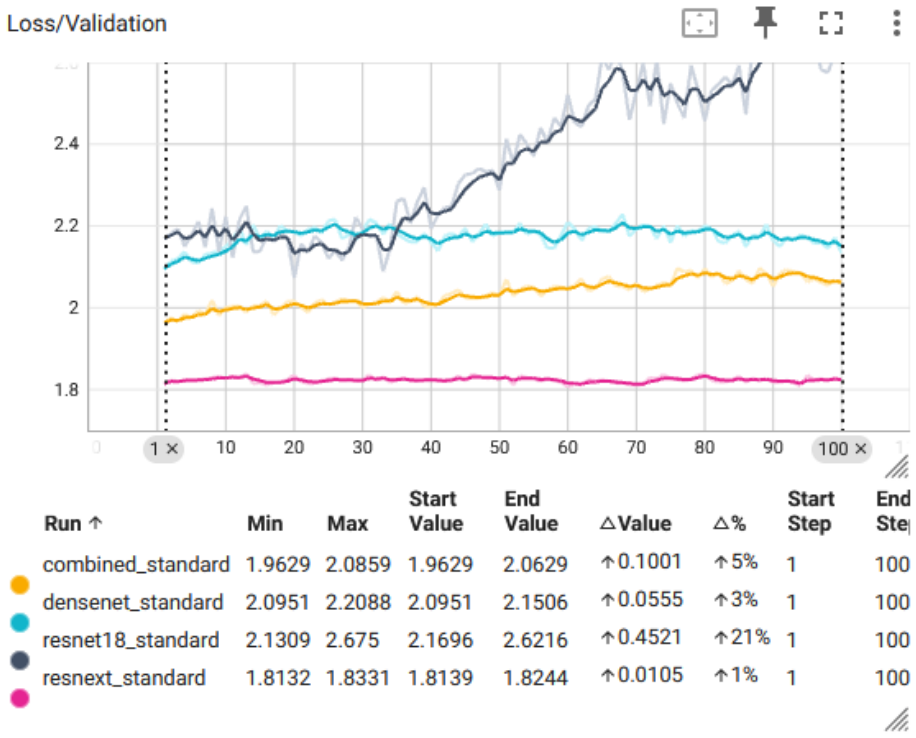
Fig. 5. A plot showcasing the training loss for the student model trained by the ResNet18, DenseNet201, ResNeXt64, Combined (DenseNet201 for US, DenseNet201 for Armenia and ResNeXt64 for Italy) ensembles using Knowledge Distillation on the US dataset.

issues in our experiment, with data imbalance causing performance loss even in the individual model training.

In Figure 9 we can see that the training loss for the Resnet18 homogeneous ensemble fluctuates between 2.17 and 2.23, averaging around 2.20. Validation loss shows significant variation between 2.15 and 2.30, with an average of about 2.22. Training accuracy stabilizes at around 0.10, while validation accuracy varies but averages close to 0.08.

We believe artificially increasing the dataset through transformation could significantly improve results. This could be used alongside a GAN to procure a similar output to the datasets we used to provide even more data and to achieve a balance between the classes. This could result in at least a more promising performance while protecting the privacy of the existent dataset.

We could also use this approach for the teacher models, to further improve their performance, as they play a key role in training the student model on the noisy dataset. This could result in better training and further improvements to student model performance.



Run	Min	Max	Start Value	End Value	△Value	△%	Start Step	End Ste
combined_standard	1.9629	2.0859	1.9629	2.0629	↑0.1001	↑5%	1	100
densenet_standard	2.0951	2.2088	2.0951	2.1506	↑0.0555	↑3%	1	100
resnet18_standard	2.1309	2.675	2.1696	2.6216	↑0.4521	↑21%	1	100
resnext_standard	1.8132	1.8331	1.8139	1.8244	↑0.0105	↑1%	1	100

Fig. 6. A plot showcasing the training validation for the student model trained by the ResNet18, DenseNet201, ResNeXt64, Combined (DenseNet201 for US, DenseNet201 for Armenia and ResNeXt64 for Italy) ensembles using Knowledge Distillation on the US dataset.

Table 3. A summary of performance for the trained ResNet18 student model evaluation. The combined model refers to DenseNet201 for Teachers 1 and 2 and ResNeXt64 for Teacher 3.

Ensemble	Implementation	Val Acc	Val Loss	Tr Acc	Tr Loss	Tr Time
ResNet18	Knowledge Distillation	0.2569	2.2012	–	1.5170	5.139 hr
ResNet18	PATE	0.2367	1.7080	0.3530	1.5215	3.628 hr
ResNeXt64	Knowledge Distillation	0.2847	1.8195	–	1.5485	2.305 hr
ResNeXt64	PATE	0.2384	1.7551	0.3209	1.6221	3.656 hr
DenseNet201	Knowledge Distillation	0.3056	2.1468	–	1.5465	2.251 hr
DenseNet201	PATE	0.3183	1.6513	0.3262	1.6027	3.637 hr
Combined	Knowledge Distillation	0.22569	1.40849	–	2.00693	2.156 hr
Combined	PATE	0.30159	1.63494	0.32394	1.58359	3.545 hr

As a final note, we provide all the resulting run logs, the best-performing teacher models used in the ensemble experiments and graphs to be checked in more detail in the GitHub repository[2] where the code for all the implementations is also posted.

[2] https://github.com/ovybe/paterockartsota/

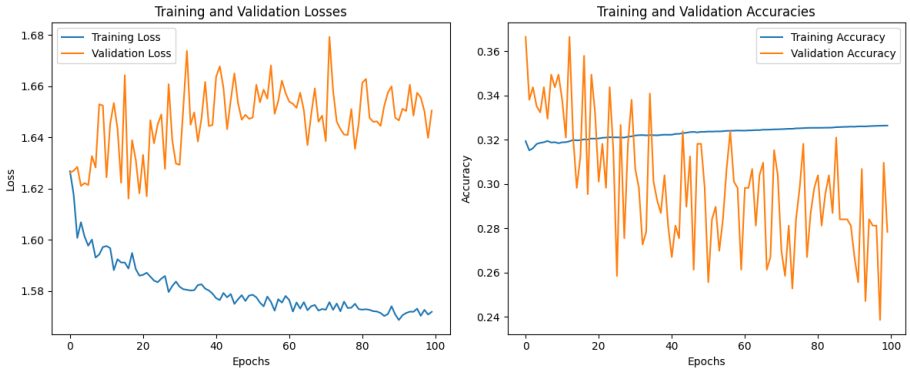


Fig. 7. A plot showcasing the training and validation losses and accuracy for the student model trained by the Combined Ensemble (DenseNet201 for US, DenseNet201 for Armenia and ResNeXt64 for Italy) using PATE on the US dataset.

7 Conclusion

In conclusion, we showcased our current research progress and gained insights on the current state of the art when it comes to the task of rock art classification. We've also showcased the existent limitations of our approach aiming to make use of the PATE framework and its alternative Knowledge Distillation for their privacy guarantee, which allows the use of sensitive data in training.

We believe the main reason for the low achieved results to be caused by the small size of a dataset featuring classes that can vary from simple to complex. This, alongside the classes differing depending on country of origin, results in the student model under-performing in comparison to the individual teacher models, which are each trained and validated in only one country.

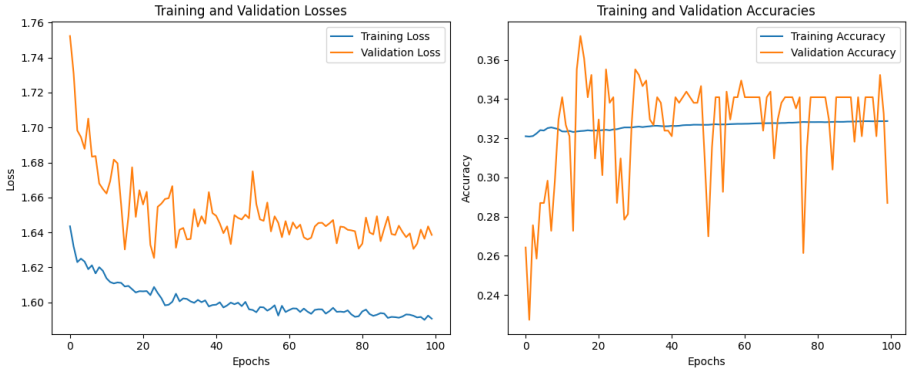


Fig. 8. A plot showcasing the training and validation losses and accuracy for the student model trained by the DenseNet201 Ensemble on the US dataset.

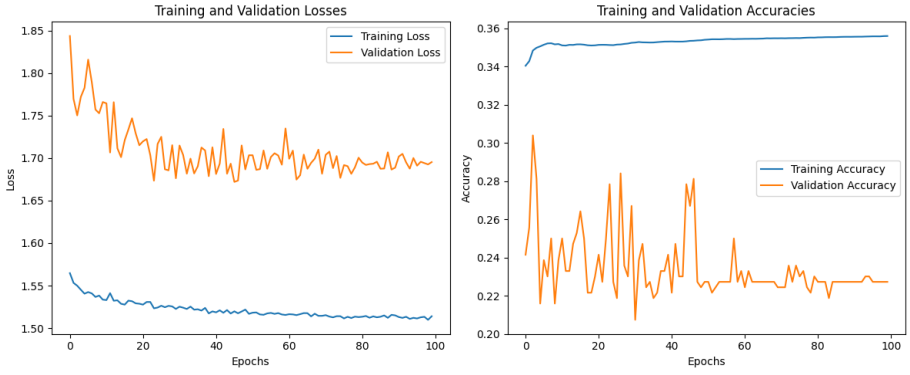


Fig. 9. A plot showcasing the training and validation losses and accuracy for the student model trained by the Resnet18 Ensemble using PATE on the US dataset.

A good possible solution to these problems would be artificially increasing the dataset size through further transformations and GAN. These require more complex research and experimenting. We use both rock art paintings and carvings, which may be differently affected by more varied and complex transformations.

The same issue could arise with GANs when it comes to classes, as we would have to generate data based on the existent sets. Adding the generated data to the training would likely improve the models' performance. Unfortunately, as it is synthetic, the resulting images could be unrealistic and cause bias, as we lack a large enough dataset to assure the robustness of the artificial output.

Even so, we believe using a voting system mechanism formed by an ensemble of models, whether used to train a student model or for data prediction is a promising path to continue researching for the task of rock art classification. This is due to allowing the mitigation of some of the most challenging aspects of rock art datasets. One such aspect is its heterogeneous data based on the culture of the region of origin.

The comparison between the two approaches also showed that PATE is the better choice for future research involving training a student model using a teacher ensemble, offering overall better performance than Knowledge Distillation with the same setup and parameters. Even so, at present, the best performance seems to stem from an individual CNN being continuously trained on data as can be concluded from our overall experiment results. This comes at the cost of being unable to train the model using sensitive data.

Another approach could be to instead simply guarantee privacy before providing said data to a student model, as the teacher models seem to instead cause an overall performance loss compared to their individual achieved performance. But this requires further research, and we believe would come with its own issues.

Our next steps are to use generated images using GANs to increase the size of the datasets and possibly ameliorate the student model underfitting, improving its performance.

We also plan to redo this experiment with a variant of the dataset where we remove the classes that contain the least amount of data and may be too similar to other classes. We believe these are causing some of the drops in performance during evaluation and their removal could significantly increase both the teacher and student model performance.

Unfortunately, this could not be done for the current experiment, but the achieved results remarkably helped point out some of the areas we can focus on improving further.

References

1. Aristizábal, A.: Making PATE Bidirectionally Private (2019), https://github.com/aristizabal95/Making-PATE-Bidirectionally-Private
2. Assiri, A.S., Nazir, S., Velastin, S.A.: Breast tumor classification using an ensemble machine learning method. Journal of Imaging **6**(6) (2020). https://doi.org/10.3390/jimaging6060039, https://www.mdpi.com/2313-433X/6/6/39

3. Bednarik, R.G.: Rock art science. Brepols (2001)
4. Boenisch, F., Mühl, C., Rinberg, R., Ihrig, J., Dziedzic, A.: Individualized pate: Differentially private machine learning with individual privacy guarantees (2022)
5. Domingo Sanz, I., May, S., Smith, C.: Communicating through rock art : an ethnoarchaeological perspective (02 2016). https://doi.org/10.13140/RG.2.1.4539.6244
6. Giuffrè, M., Shung, D.L.: Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. NPJ digital medicine **6**(1), 186 (2023)
7. Horn, C., Ivarsson, O., Lindhé, C., Potter, R., Green, A., Ling, J.: Artificial intelligence, 3d documentation, and rock art—approaching and reflecting on the automation of identification and classification of rock art images. Journal of Archaeological Method and Theory **29** (03 2022). https://doi.org/10.1007/s10816-021-09518-6
8. Jalandoni, A., Zhang, Y., Zaidi, N.A.: On the use of machine learning methods in rock art research with application to automatic painted rock art identification. Journal of Archaeological Science **144**, 105629 (2022). https://doi.org/https://doi.org/10.1016/j.jas.2022.105629, https://www.sciencedirect.com/science/article/pii/S0305440322000875
9. Kamath, H.: PATE Example (2019), https://github.com/kamathhrishi/PATE
10. Kang, J., Ullah, Z., Gwak, J.: Mri-based brain tumor classification using ensemble of deep features and machine learning classifiers. Sensors **21**(6) (2021). https://doi.org/10.3390/s21062222, https://www.mdpi.com/1424-8220/21/6/2222
11. Lyu, L., Chen, C.H.: Differentially private knowledge distillation for mobile analytics. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1809–1812 (2020)
12. Papernot, N., Abadi, M., Úlfar Erlingsson, Goodfellow, I., Talwar, K.: Semi-supervised knowledge transfer for deep learning from private training data (2017)
13. Pratiwi, R., Nurmaini, S., rini, d., Naufal Rachmatullah, M., Darmawahyuni, A.: Deep ensemble learning for skin lesions classification with convolutional neural network. IAES International Journal of Artificial Intelligence (IJ-AI) **10**, 563–570 (09 2021). https://doi.org/10.11591/ijai.v10.i3.pp563-570
14. Tsang, R., Brady, L.M., Katuk, S., Taçon, P.S., Ricaut, F.X., Leavesley, M.G.: Agency, affect and archaeologists: Transforming place with rock art in auwim, upper karawariarafundi region, east sepik, papua new guinea. Rock Art Research: The Journal of the Australian Rock Art Research Association (AURA) **38**(2), 183–194 (2021)
15. Wang, J., Bao, W., Sun, L., Zhu, X., Cao, B., Philip, S.Y.: Private model compression via knowledge distillation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 1190–1197 (2019)