Rock Art Classification Through Privacy-Guaranteed Ensemble Machine Learning

Ovidiu-Emilian Buțiu
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 underperforming with an average accuracy of around 30-35% and that the best approach to rock art classification to be individual CNN architectures, which have achieved an accuracy between 70-80% We believe the main cause is the small, unbalanced and very heterogeneous dataset, which could be further artificially increased in future research through GANs and transformations.**

***Index Terms*—Rock Art, Classification, ResNeXt64, Resnet18, DenseNet201, Voting Mechanism Ensemble, ML, PATE, Knowledge Distillation.**

I. 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 [1]. They also vary in visual depiction due to cultural differences [2], outsider and environmental damage [1], 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:

- Can 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.

II. Problem Description

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 [3], 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 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) [4]. 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.

III. Related Work

According to Robert Bednarick [5], 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 **??**, which instead depicts different techniques, including carving, used to display objects and animals.

A. Methodology

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

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 provided by the supervisor in the countries US, Spain, Italy+Bulgaria+Spain, 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.

TABLE I
STATISTICS OF THE PRIVATE ROCK ART USED DATA

Country	*Count*	*Train Split*	*Test Split*	*Person*	*Goat*	*Stag*	*Circle*	*Spiral*	*Zigzag*	*Cross*
Armenia	106	95	11	91	132	18	53	9	27	10
Italy+Spain+Bulgaria	66	59	7	134	82	28	84	57	44	19
US	91	82	9	108	44	4	104	58	66	31
Total	263	236	27	333	258	50	241	124	137	60

B. Dataset Difficulties

The data, as seen in Table I, 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 biased towards. Similarly, the "Person" label may suffer from a sbias issue in comparison to some of the other labels having similar characteristics. One such example would be the horns being mistaken as the spears held by some of the painted humans. Or the heads of figures being mistaken for circles or spirals. We aim to use Generative Adversarial Networks (GANs) to ameliorate this possible bias by using existent data as input for new artificially generated data.

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).

C. State of the Art

When it comes to the state of the art for rock art classification, there is a distinct lack of primary literature compared to other research topics. In our search through Web of Science, we found under 20 total research papers close to our task.

Filtering through these papers to only contain related work to automated rock art classification resulted in an even smaller number of $\approx$ 11 research papers. All of these papers were primary literature, with no secondary literature to be found for our topic.

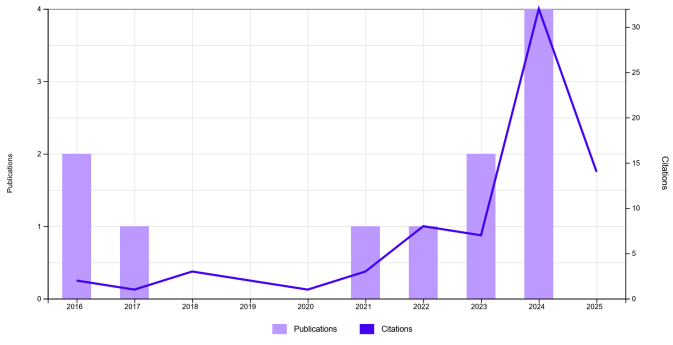

Fig. 1. Showcase of the amount, publication date and total yearly citations of found papers close to the topic of rock art classification through a search of "rock art classification ml" using Web of Science.

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 unfortunately used datasets that can only be considered similar in characteristics to the rock art datasets 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 I, 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. [6] [7] [8] 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.

Existing works primarily focus on single-region rock art detection or classification, yielding strong local results but lackluster performance in generalization. Even advanced approaches like Faster R-CNN [3] struggle with complex motifs (e.g., "animal" or "human"), achieving only 32.47% mAP, though simpler classes like "boat" reached 60% precision. This performance gap is argued to be stemming from representation variability and limited training data.

Recent Australian research [9] achieved 80-90% accuracy in rock art detection using VGG/Inception/ResNet, but motif classification remains challenging. While these studies demonstrate promising single-country results, they neglect two critical issues we address: (1) cross-cultural motif variability and (2) data privacy in collaborative research.

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.

Rock art classification faces two key challenges: regional motif variations causing overfitting, and privacy concerns preventing data sharing. We drew inspiration from medical imaging, where similar challenges exist - heterogeneous data (varying shapes/complexity) and strict privacy requirements.

Recent medical approaches demonstrate effective ensemble methods. One paper [6] combined InceptionV3, Inception-ResNetV2, and DenseNet201 for skin lesion classification, achieving 97.23% accuracy. Another research [7] used voting among top classifiers for breast tumors (99.42% accuracy). Another paper [8] employed feature ensemble from CNNs with SVM-RBF for brain tumors (93-98% accuracy), showing transfer learning's potential for complex data.

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. [4] 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.

Unlike existing approaches, we employ the PATE framework, which enables distributed training of teacher models on private datasets across different systems. Through a noise-injected voting mechanism, these teachers collaboratively train a student model while preserving data confidentiality. This scalable design allows seamless integration of new models to compensate for individual weaknesses, using only queries where teacher consensus exceeds threshold T for student training.

The framework enables secure use of sensitive rock art data while addressing dataset limitations. Standard PATE sufficiently validates our ensemble approach's viability, though future work could explore individualized privacy [10] if varying privacy requirements emerge. While offering improvements, these implementations pose evaluation challenges beyond our current scope.

An alternative to PATE is knowledge distillation, where a compact student model learns from either a large teacher model or an ensemble. Unlike PATE's voting system, this approach focuses on soft labels (probability vectors) that capture nuanced class relationships, enabling comparable accuracy with lower computational costs [11]. While one proposed hybrid method combining PATE with distillation [11] exceeds our scope, the core concept offers a viable alternative that may outperform PATE, when it comes to smaller datasets, through more efficient training.

For private knowledge distillation, we adapt a batch-based noise injection approach [12], simplifying it to operate sample-wise. Each ensemble teacher's outputs are perturbed individually after adaptive norm clipping, preserving confidentiality while maintaining the distillation benefits. This method balances privacy protection with the computational efficiency advantages that make knowledge distillation attractive for resource-constrained applications.

Synthetic data generation offers an alternative solution to data scarcity and privacy concerns. Modern techniques like Generative AI Networks (GANs) and Variational Autoencoders (VAEs) can produce realistic synthetic datasets that serve as privacy-preserving proxies for real data, as demonstrated in healthcare applications [13]. While synthetic rock art data could theoretically enable unrestricted ensemble methods by providing inherent privacy guarantees, this approach proves inefficient given our already limited dataset. Instead, we augment our existing data with GAN-generated samples to potentially improve model performance and address class imbalance in both PATE and knowledge distillation.

However, synthetic data introduces risks including amplified bias, reduced interpretability, and artificial correlations [13]. Generated images may perpetuate existing imbalances, potentially degrading model performance and creating "blind spots" for underrepresented classes. We focus on mitigating these risks through careful documentation and bias monitoring, focusing on controlled augmentation rather than full dataset replacement for our experiments.

For the task of rock art classification, we chose to implement the standard PATE framework and compare it with an implementation of a similar but less complex ensemble approach 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.

IV. Contributions

Unfortunately, automated rock art classification is currently still a niche topic, which offers very little documented research compared to other topics. This results in a limited amount of information to work with, based mainly on the few already existent papers as close to our task as possible, whether from the domain of computer science, or archaeological papers featuring manual classification of rock art data.

Thus, the focus of our research paper is to provide a well-performing ML approach to the task of rock art classification. We test the potential of ensemble approaches formed from models specialized on specific regions through the standard approach of Private Aggregation of Teacher Ensembles, as that may provide the possibility of researching privacy guarantees for sensitive rock art data in the future.

To combat the limited information we have encountered during our search, we also provide an in-depth analysis of the dataset and the achieved results by both the individual models used for the ensemble framework and the ensemble solution used for our task.

We also experiment with generative images and document the effects of adding synthetic data to the existent used rock art data with the goal of further improving performance and combating the issue of data imbalance caused by the limited availability of rock art data.

Based on the state of the art presented in the related work, we notice a large research gap in the use of an ensemble approach for the task of rock art classification, which resulted in the need of researching different topics that use datasets with similar characteristics to the ones derived from procured rock art data.

Most of the research papers that actually focus on the task of rock art classification mainly use CNN models that offer great results in the task of classification for rock art data provided for one region. In short, our objectives and contribution to the topic of rock art classification are the following:

- Benchmark CNN performance (ResNet18, ResNeXt101, DenseNet201) across regional datasets.
- Evaluate ensemble methods viability (PATE and Knowledge Distillation) for motif classification.
- Analyze classification results across diverse rock art datasets.
- Assess model performance with/without GAN-generated images.

V. Solution

A. PATE: Private Aggregation of Teacher Ensembles

Based on the described requirements for the domain that can be perceived from the previous sections, data privacy, and ensemble machine learning can play a big role in aiding the archaeologists when it comes to the classification of rock art data, as they allow the possibility of utilizing private data for categorization and building a scalable ensemble of models to boost the overall accuracy. The ensemble approach chosen to tackle these two needs is *Private Aggregation of Teacher Ensembles*, or PATE for short.

A paper on scalable private learning [14] introduces PATE as an ensemble of teacher models trained on disjoint data partitions that transfers knowledge to a student model while preserving privacy. This approach allows flexible model selection and creates diverse predictions. The framework allows the aggregation of teacher votes through a differentially private mechanism: adding Laplacian noise to vote counts before selecting the majority class. [14]

The final part of the framework is the student model, which is trained by knowledge transfer from the ensemble of teacher models using public unlabeled data. To limit the privacy cost stemming from labeling, the student model is trained in a semi-supervised way; by employing a specified amount of queries, which are constructed for a subset of the public data, to the aggregation mechanism mentioned previously, the privacy cost is fixed. At the same time, it diminishes the value of attacks that risk salvaging training data by analyzing model parameters. The student, ultimately, only sees the shared data and privacy-guaranteeing labels.

B. Used Individual Models

The provided ensemble solution uses the following CNN architectures in the attempted and thoroughly documented experiments:

- **ResNet-18:** Introduces residual learning with skip connections to address vanishing gradients in deep networks, using 18 layers of basic residual blocks [15].
- **ResNeXt-101 (64×4d):** An extension of ResNet through the employment of grouped convolutions (64 groups with 4D bottlenecks) to enhance parameter efficiency while maintaining 101-layer depth [16].
- **DenseNet-201:** Leverages dense connectivity, where each layer connects to all subsequent layers, promoting feature reuse and yielding high accuracy with 201 layers despite higher memory overhead [17].

This is because they were the closest available choices provided by our used technologies to the similar work using ensemble ML in the task of classification.

VI. Experimental Results

A. Research

Before our experiment, we reviewed a PyTorch-based PATE implementation from a GitHub repository [18], inspired by a research paper demonstrating its strong privacy guarantees for training data [4]. While the original implementation focused on MNIST image classification, we adapted it by replacing the model with our own and creating a custom class compatible with our datasets. Due to the need of compatibility with Syft, the library providing evaluation for PATE's black-box implementation, we restricted our framework to PyTorch.

B. Setup

Our experiment's code is available on GitHub.[1] We began by comparing model performance using Python 3.7 in an Anaconda environment, requiring older package versions (PyTorch 1.1.0, TorchVision 0.3.0, cudatoolkit 9.0, protobuf 3.20.1, matplotlib, and Pillow (< 7).) due to Syft's PATE implementation being unavailable in newer releases [19]. The setup ran on an RTX 3060 Ti (4,864 CUDA cores), an Intel® Core™ i9-11900F, and 32GB DDR4 RAM.

We trained each model five times per dataset on newer PyTorch for efficiency, then converted them to older versions for PATE compatibility. The best-performing models from these 15 runs were selected for ensemble testing—first with identical models, then with a mixed ensemble of top US, Armenia, and Italy+Spain+Bulgaria models. For the student model, we used ResNet18 (which was fastest at ≈ 3.7 hours/100 epochs) and trained it on the US dataset with added Laplace noise (PATE) or noised teacher logits (Knowledge Distillation).

This approach allowed direct comparison of student model performance across different teacher ensembles while maintaining privacy constraints. The mixed ensemble aimed to leverage the strengths of the best performing models for improved generalization.

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 normalize 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.

TABLE II
COMPARISON OF VALIDATION RESULTS FOR THE BEST PERFORMING MODELS ON THE STANDARD DATASETS.

Model	Dataset	Val Acc	Val Loss	Recall Avg	F1 Score	Tr Time (h)
DenseNet	US	0.87	0.80	0.68	0.70	3.12
ResNet	US	0.78	1.17	0.57	0.58	0.87
ResNext	US	0.82	0.80	0.64	0.65	6.80
DenseNet	Armenia	0.86	0.59	0.84	0.84	3.13
ResNet	Armenia	0.84	0.66	0.78	0.79	1.00
ResNext	Armenia	0.86	0.61	0.80	0.82	6.04
DenseNet	Italy+Bulgaria+Spain	0.77	1.11	0.70	0.71	3.22
ResNet	Italy+Bulgaria+Spain	0.75	1.44	0.68	0.70	1.07
ResNext	Italy+Bulgaria+Spain	0.83	0.97	0.73	0.75	6.30

TABLE III
PERFORMANCE SUMMARY FOR THE RESNET18 STUDENT MODEL. THE COMBINED MODEL USES THE BEST TEACHER MODELS (DENSENET201 FOR TEACHERS 1/2, RESNEXT64 FOR TEACHER 3).

Ensemble	Method	Val Acc	Val Loss	Tr Acc	Tr Loss	Tr Time
ResNet18	KD	0.2569	2.2012	–	1.5170	5.14 h
ResNet18	PATE	0.2367	1.7080	0.3530	1.5215	3.63 h
ResNeXt64	KD	0.2847	1.8195	–	1.5485	2.30 h
ResNeXt64	PATE	0.2384	1.7551	0.3209	1.6221	3.66 h
DenseNet201	KD	0.3056	2.1468	–	1.5465	2.25 h
DenseNet201	PATE	0.3183	1.6513	0.3262	1.6027	3.64 h
Combined	KD	0.2257	1.4085	–	2.0069	2.16 h
Combined	PATE	0.3016	1.6349	0.3239	1.5836	3.55 h

C. Achieved Results

The training was done using PyTorch version 2.5.1, with similar data pre-processing 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 II and student model performance at Table III. We also provide summarized comparison graphs for the resulting student model validation accuracy for both implementations in Figure 2. We also provide a comparison graph for the F1 Score achieved by the ensembles evaluated for PATE in Figure 3.

The performance summary of these models can be reviewed at Table II. 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.

The model trained on the US dataset (Teacher 1) struggled most with crosses and slightly with spirals and stags (F1: $\approx$ 0.6–0.7 vs. overall average $>$ 0.8), likely due to the absence of zigzag test data, which resulted in a score of 0. The Armenia-trained model (Teacher 2) had difficulty identifying stags, possibly because of dataset imbalance, such as 132 goat compared to 18 stag motifs, leading to a bias toward goats, though other classes performed well (F1: 0.7–0.9). The Italy+Bulgaria+Spain-trained model (Teacher 3), with the smallest dataset, performed poorly on crosses, stags, and zigzags, likely due to extreme class imbalance (e.g. only 4 stag motifs vs. 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 rock art sites in Italy+Bulgaria+Spain 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 a F1 score of $\approx 0.6-0.9$.

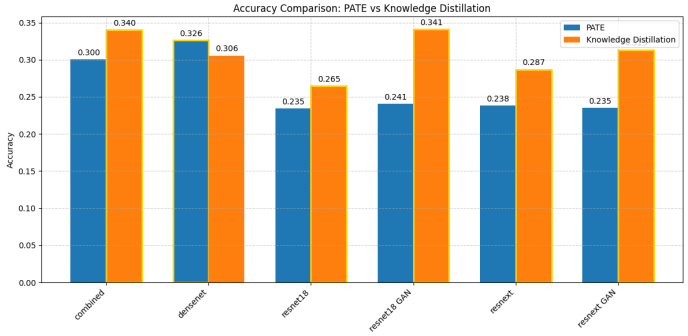

Fig. 2. A plot showcasing a validation accuracy result comparison between the PATE and Knowledge Distillation ensemble frameworks after training each ensemble for 100 epochs.

For the ResNeXt64 homogeneous teacher ensemble individual models, we observed an average accuracy of ≈ 0.23–0.24 with an average loss between 1.60 and 1.82. The F1 score average ranged between 0.19 and 0.21, indicating significant challenges in detecting certain classes. While most classes achieved moderate performance (peaking at F1 ≈ 0.29), the models consistently struggled with crosses, stags, and zigzags (F1: 0.14–0.23), requiring approximately 3 hours per 100 epochs of training time.

Among all architectures, DenseNet201 demonstrated the strongest results (accuracy: 0.29–0.33, F1: 0.23–0.29) despite its longer training requirements ($\approx$6 hours/100 epochs). Cross motifs proved most problematic (F1: ≈ 0.14), followed closely by stag motifs (F1: ≈ 0.19).

The persistent gap between training and validation performance suggests notable overfitting, shown by dataset imbalances that particularly affected these challenging motifs. While acquiring additional rock art data remains difficult, potential mitigation strategies include further optimizing the model and enhanced pre-processing techniques to improve feature differentiation.

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

For Knowledge Distillation implementations, the ResNet18 ensemble-trained student model showed the weakest performance, achieving a validation accuracy average of 0.265 as seen in Figure 2 with an average validation loss of 2.201. The ResNeXt64 ensemble demonstrated modest improvement, reaching a validation accuracy average of 0.287 with the best average validation loss of 1.820. The DenseNet201 ensemble delivered the strongest accuracy performance at 0.306 out of the models trained on the standard datasets, though with a slightly higher average validation loss of 2.147.

In PATE implementations, the ResNet18 ensemble again showed the lowest performance with an average validation accuracy of 0.235 and average loss of 1.708, while ResNeXt64 offered minimal gains with an average accuracy of 0.238 and average loss of 1.755). DenseNet201 maintained its lead with the best accuracy (0.326) as shown in Figure 2 and lowest validation loss (1.651).

Our mixed ensemble approach, combining the best-performing teacher models (DenseNet201 for US and Armenia datasets, ResNeXt64 for Italy+Bulgaria+Spain) yielded interesting results. In Knowledge Distillation, this configuration achieved comparable performance to the ResNet18 GAN homogeneous ensemble with a validation accuracy of 0.340 (F1: 0.197) and validation loss of 1.820.

For PATE, the student model showed slightly lower performance relative to the DenseNet201 homogeneous ensemble, reaching an average validation accuracy of 0.300 (F1: 0.229) with an improved validation loss of 1.651. While the accuracy gains were modest, the F1 scores demonstrated more consistent improvement across all motif classes, putting the mixed model on top, as seen in Figure 3.

As expected, Knowledge Distillation outperformed PATE for the same teacher ensemble, likely due to the small dataset size. However, both methods showed limited accuracy (0.2–0.3), suggesting student model underfitting. This is evidenced by similar training/validation metrics with fluctuating validation results. The primary limitation appears to be dataset quality, with both size and imbalance causing lackluster performance.

To ameliorate this issue, we attempted dataset augmentation using GANs, generating ∼200 artificial training images per country. While this slightly reduced overfitting in teacher models (with minor performance trade-offs), it unfortunately provided only a small improvement for the PATE student model's underfitting. The augmented ensemble showed close accuracy ranges (0.3–0.35) to the original.

We also experimented with merging "Goat" and "Stag" under a more general "Animal" label, in an attempt to solve some of the data imbalance, as "Stag" seems to offer the lowest amount of data in every dataset. This resulted in slight increases in performance for the teacher models' training for the datasets, and low to no increase in the resulting student model's training for PATE.

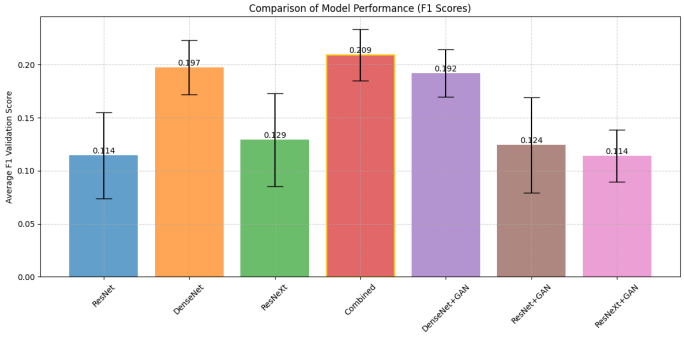

Fig. 3. A comparative column graph showcasing the F1 score performance of the PATE models.

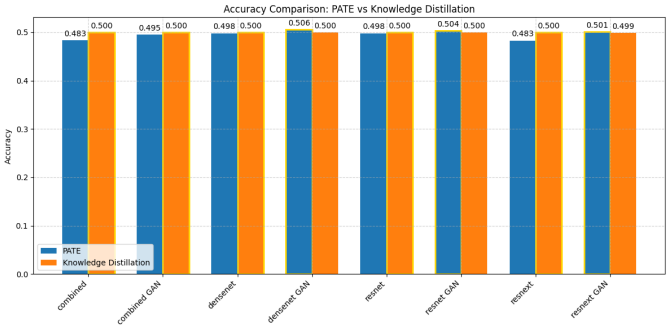

Fig. 4. A plot showcasing a validation accuracy result comparison between the PATE and Knowledge Distillation ensemble frameworks using the trained teacher models on the dataset containing only the most populated labels "Person" and "Circle". Each ensemble was trained for 100 epochs.

TABLE IV
COMPARISON OF MODELS TRAINED ACROSS DATASETS CONTAINING ONLY THE MOST POPULATED LABELS PERSON AND CIRCLE+SPIRAL.

Model	Dataset	Val Acc	Val Loss	Recall Avg	F1 Score	Tr Time (h)
DenseNet	US	0.98	0.05	0.98	0.98	1.92
ResNet	US	0.98	0.05	0.98	0.98	0.32
ResNext	US	0.95	0.12	0.95	0.95	3.84
DenseNet	Armenia	0.99	0.02	0.99	0.99	1.91
ResNet	Armenia	0.99	0.04	0.99	0.99	0.31
ResNext	Armenia	0.98	0.04	0.98	0.98	3.82
DenseNet	Italy+Bulgaria+Spain	0.93	0.19	0.93	0.93	2.47
ResNet	Italy+Bulgaria+Spain	0.91	0.35	0.90	0.90	0.41
ResNext	Italy+Bulgaria+Spain	0.93	0.20	0.93	0.93	5.10

To assess model performance on the most populated labels, a variant of each dataset was created containing only the two most frequent classes, as seen in Table I: "Person" with 333 motifs and a generalized "Circle" class with 365 motifs which merges the original "Circle" and "Spiral" labels due to their similarity.

When it came to PATE and KD's performance on these Person+Circle datasets, a significant increase can be noted in Figure 4, where all the student models are showing close to 0.5 accuracy. The addition of generated images, offers a slight increase of performance, but the values still hover close to 0.5 accuracy.

As shown in Table IV, DenseNet201 achieved the highest overall accuracy on these two-class datasets, though it required the longest training time (≈ 2-–3.5 hours). Other models trained faster but with slightly lower accuracy.

TABLE V
COMPARISON OF GAN MODELS TRAINED ACROSS DATASETS CONTAINING ONLY THE MOST POPULATED LABELS PERSON AND CIRCLE+SPIRAL.

Model	Dataset	Val Acc	Val Loss	Recall Avg	F1 Score	Tr Time (h)
DenseNet	US	0.97	0.07	0.97	0.97	18.50
ResNet	US	0.95	0.13	0.95	0.95	2.00
ResNext	US	0.95	0.14	0.95	0.95	30.64
DenseNet	Armenia	0.99	0.03	0.99	0.99	17.50
ResNet	Armenia	0.99	0.03	0.99	0.99	1.80
ResNext	Armenia	0.99	0.02	0.99	0.99	20.00
DenseNet	Italy+Bulgaria+Spain	0.92	0.26	0.92	0.92	8.40
ResNet	Italy+Bulgaria+Spain	0.89	0.47	0.89	0.89	2.34
ResNext	Italy+Bulgaria+Spain	0.93	0.23	0.93	0.93	17.44

Adding artificially generated motifs produced results similar to the merged and standard datasets, as seen in Table V. ResNeXt64 benefited the most, even surpassing DenseNet in the Armenia and Italy+Bulgaria+Spain datasets. DenseNet201's performance slightly declined with generated data, while ResNet18 showed mixed results—small gains in Armenia but slight decreases elsewhere.

We believe based on all the achieved experimental results that focusing on a simpler approach, improving on the individual teacher performance by further tweaking the parameters and improving the overall implementation would be the recommended choice. Using artificial images along the original data had an effect in ameliorating the overfitting issues suffered by the models, but further tweaking the models and experimenting further with transformations to be the next step in improving performance.

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.

VII. CONCLUSIONS

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.

We thought a good possible solution to these problems would be artificially increasing the dataset size through images provided by a GAN model. Experimenting with this approach resulted in ameliorating the existent overfitting issues suffered by the teacher models, at the cost of performance. Unfortunately, this approach had no effect on the student model training, achieving the same low performance as before.

The comparison between two approaches, PATE and Knowledge Distillation, using the same setup, showed that a simpler approach to this problem may be preferred, as both frameworks, albeit similar, suffered from the same underfitting issues. This resulted in PATE and Knowledge Distillation offering below average results, while the teacher models, which use well-known CNN architectures, offer better overall scores.

Thus, 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.

VIII. FUTURE WORK

Based on the achieved results, we believe the best approach would be to avoid the complex ensemble frameworks and instead work with individual models, focusing on improving overall scores and further combating overfitting.

Ensemble models seem to unfortunately have problems when it comes to generalization of the data, achieving underwhelming results in comparison to the CNN architectures trained to be teacher models. This could mean that a better approach to the task of rock art classification is to improve individual model performance, focusing on either a specialized model for each country or region, or a more general model trained using all available rock art data, depending on achieved performance.

Next steps would be improving model performance as previously mentioned and finding more ways to combat the models overfitting through further researching with various parameter setups and implementation techniques. Another possible path to take would be improving the performance through the data used for training the model.

This would entail various experiments with data preprocessing through various well-known techniques to aid the model in better understanding the patterns to look for during training.

If still pursuing the use of 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.

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

REFERENCES

[1] R. Tsang, L. M. Brady, S. Katuk, P. S. Taçon, F.-X. Ricaut, and M. G. Leavesley, "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)*, vol. 38, no. 2, pp. 183–194, 2021.

[2] I. Domingo Sanz, S. May, and C. Smith, *Communicating through rock art : an ethnoarchaeological perspective*, 02 2016.

[3] C. Horn, O. Ivarsson, C. Lindhé, R. Potter, A. Green, and J. Ling, "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*, vol. 29, 03 2022.

[4] N. Papernot, M. Abadi, Úlfar Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," 2017.

[5] R. G. Bednarik, *Rock art science*. Brepols, 2001.

[6] R. Pratiwi, S. Nurmaini, d. rini, M. Naufal Rachmatullah, and A. Darmawahyuni, "Deep ensemble learning for skin lesions classification with convolutional neural network," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, pp. 563–570, 09 2021.

[7] A. S. Assiri, S. Nazir, and S. A. Velastin, "Breast tumor classification using an ensemble machine learning method," *Journal of Imaging*, vol. 6, no. 6, 2020. [Online]. Available: https://www.mdpi.com/2313-433X/6/6/39

[8] J. Kang, Z. Ullah, and J. Gwak, "Mri-based brain tumor classification using ensemble of deep features and machine learning classifiers," *Sensors*, vol. 21, no. 6, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/6/2222

[9] A. Jalandoni, Y. Zhang, and N. A. Zaidi, "On the use of machine learning methods in rock art research with application to automatic painted rock art identification," *Journal of Archaeological Science*, vol. 144, p. 105629, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0305440322000875

[10] F. Boenisch, C. Mühl, R. Rinberg, J. Ihrig, and A. Dziedzic, "Individualized pate: Differentially private machine learning with individual privacy guarantees," 2022.

[11] L. Lyu and C.-H. Chen, "Differentially private knowledge distillation for mobile analytics," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 1809–1812.

[12] J. Wang, W. Bao, L. Sun, X. Zhu, B. Cao, and S. Y. Philip, "Private model compression via knowledge distillation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 1190–1197.

[13] M. Giuffrè and D. L. Shung, "Harnessing the power of synthetic data in healthcare: innovation, application, and privacy," *NPJ digital medicine*, vol. 6, no. 1, p. 186, 2023.

[14] N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, and Úlfar Erlingsson, "Scalable private learning with pate," 2018.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[16] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *CVPR*, 2017.

[17] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017.

[18] H. Kamath. (2019) PATE Example. [Online]. Available: https://github.com/kamathhrishi/PATE

[19] A. Aristizábal. (2019) Making PATE Bidirectionally Private. [Online]. Available: https://github.com/aristizabal95/Making-PATE-Bidirectionally-Private