# Smart Assisted Archaeological Script Analysis System

Prof. Kavita Dige[1], Sunit Dhopte[1], Akshada Padale[1], Shreya Walde[1]
[1] *G H Raisoni College of Engineering Management, Pune.*
[1] *(An Autonomous Institute affiliated to Savitribai Phule University Pune)*

**Abstract: The Smart Assisted Archaeological Script Analysis System, relies on the innovative approaches of image processing and machine learning to automate the study of old inscriptions. Old methods of analysing an inscription consume a lot of time and effort. This system streamlines critical phases such as image enhancement, extraction, and classification using deep learning models in order to deliver faster and more accurate analyses. This way, the system will allow archaeologists to work faster and with more accuracy while saving the precious information about history. The new approach helps not only in archaeological research but also gives light toward a better understanding of ancient civilizations and their writing systems.**

**Keywords: Machine learning, deep learning, image recognition and classification, ancient script recognition, Convolutional neural network.**

## 1. INTRODUCTION

The study of ancient inscriptions is an important aspect for archaeologists to study archaeology. However, traditional methods of analysing these inscriptions are often slow, labour-intensive, and prone to human error, limiting the speed and scope of research. To address these challenges, the Smart Assisted Archaeological Script Analysis System is designed to revolutionize the field by utilizing advanced image processing and machine learning techniques. The system enhances efficiency of the research while preserving the historical integrity of the artifacts. The project’s primary objective is to create an AI-powered tool that assists archaeologists in identifying, contextualizing, and translating inscriptions, thereby improving the efficiency of artifact analysis.

## 2. LITERATURE SURVEY

Recent work in AI and machine learning has shown great impact in the field of archaeology and also in digitizing ancient scripts. This paper presents the problem of handwritten recognition of Brahmi, which forms the core of the Pali language, using OCR systems. They enhance the existing OCR techniques; apply machine learning principles on a data set that helps in achieving the accuracy of the results compared to others like Akkhara-Muni and overcome difficulties arising due to the script's ancient origins and lack of available datasets.

In the wider archaeological context, this use of AI in Artificial Intelligence in Archaeology is illustrated with AI techniques involving Convolutional Neural Networks (CNN) and Support Vector Machines (SVM), which successfully identified bone marks and satellite images with 91% and 83% accuracy, respectively.

Deep Learning in Indus Valley Script Digitization: Probing the efficacy of ASR-net-ASR is depicted as a system that digitizes Indus Valley seals with a 95 percent success rate in symbol recognition. This is simply some kind of OCR for languages, as the term goes, as if one's computer were trying to decipher something other than modern language. Digitization of the undeciphered script of the Indus civilization is one area where instrumental tools have been elevated to date. Artificial intelligence and deep learning keep changing new strategies in archaeological research into ancient tongues' preservation.

| Sr. No. | Paper Title and Year | Details of Publication | Findings from Papers |
|---|---|---|---|
| 1 | “Artificial Intelligence in The Field of Archaeology January- 2024” | University of Gothenburg Department of Historical Studies | The thesis highlights case studies where AI successfully identified previously unknown archaeological sites, showcasing its potential to revolutionize surveys. |

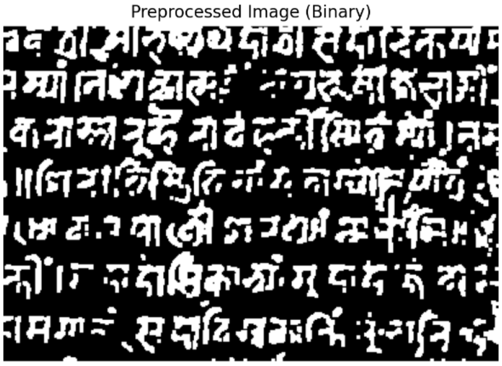
| | | | |
|---|---|---|---|
| 2 | "IArch: An AI Tool for Digging Deeper into Archaeological Data" | Conference: 2023 IEEE 35th International Conference on Tools with Artificial Intelligence | The paper is representing the experimental findings on integrating K-means clustering and Random Forest models to analyse genetic and archaeological data. |
| 3 | "Eyes of the machine: AI-assisted satellite archaeological survey in the Andes January-2024" | Cambridge University Press & Assessment | The use of AI-assisted methods has led to more comprehensive and accurate mapping of archaeological sites. |
| 4 | "Uncovering Archaeological Sites in Airborne LiDAR Data with Data-Centric Artificial Intelligence" | Conference: 2023 IEEE | The approach combines novel data processing, augmentation, and training techniques to reduce false positives in identifying burial mounds. |
| 5 | "Debating AI in Archaeology: applications, implications, and ethical considerations - 2024" | Sheffield Hallam University | Archaeology can be beneficial from AI advancements while contributing to the practical discussions of its role in society. |

## 3. METHODOLOGY

### 3.1 Data Preparation

3.1.1 Data Collection: A character image dataset from various languages, including the Indus script, etc, was collected. For each character image, there exists a subdirectory for the different types of classes, storing it as it is in the gray scale format.
3.1.2 Preprocessing: Each character image was resized to 28x28 pixels and normalized. The images were flattened in one dimension for the SVM model, whereas in their 2D form for the CNN model. Pixel values are normalized, scaling pixel values into [0, 1] range for the CNN model.
3.1.3 Label Encoding: the labels, or rather the language classes, were encoded as integers using a LabelEncoder to convert categorical labels into the numerical format for training

Preprocessed Image (Binary)



### 3.2 Model Architecture

3.2.1 Support Vector Machine (SVM)

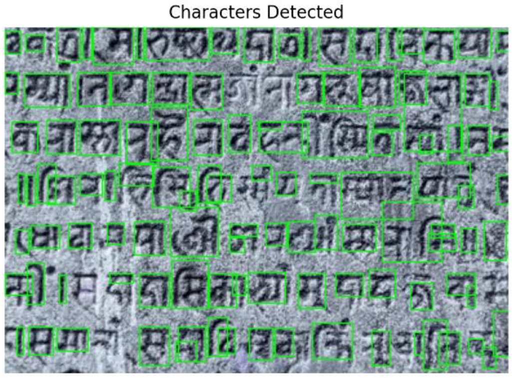
3.2.1.1 Feature Extraction: As SVM takes flattened image arrays as the input, this 28x28 pixel image was reshaped to form a 784-dimensional feature vector. That feature vector then entered into the classifier under SVM.
3.2.1.2 Training SVM classifier with a linear kernel is applied. A model comprising 80% of the data set was used for training, and 20% for testing. The train_test_split function divided the dataset.
3.2.1.3 Model Saving: The SVM model was saved for subsequent prediction runs after the training process using joblib.
3.2.2 Convolutional Neural Network (CNN)
3.2.2.1 Architecture: The CNN model comprises two 2D convolutional layers, successively followed by max-pooling layers. It had two convolutional layers with 32 and 64 filters, respectively, with a kernel size of 3x3 for each. It is followed by a fully connected layer with 128 units and finally an output layer with softmax activation for multi-class classification.
3.2.2.2 Training: The CNN model uses the adam optimizer and categorical cross entropy as the loss function. It was trained for epochs = 10 and batch size = 32.
3.2.2.3 Model Saving: after the training, it saved a model to be used, if necessary, in the future as an .h5 file.

### 3.3 Character Segmentation

3.3.1 Preprocessing for Segmentation: Input images were first converted to binary format using a thresholding technique. Otsu's thresholding method was applied to ensure robust binarization across varying image lighting conditions.

3.3.2 Contour Detection: Character segmentation, using this algorithm, relied on contour detection to locate outer bounds of individual characters in the binary image. Each bounding box was formed for every detected character.

3.3.3 Character Cropping: Every character was cropped based on the bounding box and forwarded to the classifiers for prediction.




### 3.4 Feature Extraction

3.4.1 SVM Feature Extraction: The cropped character images were resized to 28x28 pixels and flattened into a 784-dimensional feature vector for SVM classification.

3.4.2 Extract Feature of CNN: Images of characters cropped to 28x28 pixels, normalized and resized to a 28x28x1 tensor before fed the input requirement to the CNN model.

### 3.5 Categorization

3.5.1 SVM Prediction: The input of segmented characters' pre-flattened feature vectors into a pre-trained SVM model was undertaken. Output of the SVM was the class label assigned to each character, that is, the language.

3.5.2 CNN Prediction: The segmented characters are given as input as 28x28 grayscale images to the pre-trained CNN model. The output to predict the CNN model of class label based on the softmax output was the class with maximum probability.

3.6 Evaluation Metrics: The models were compared using the following metrics:

3.6.1.1 Accuracy: The proportion of correctly classified characters in the test set.

3.6.1.2 Prediction Speed: The time taken to classify characters in a given image.

3.6.1.3 Majority Prediction Accuracy: The ability of the model to correctly predict the majority language for images containing multiple characters.

### 3.7 Time Analysis for Predictive Use

3.7.1 Measurement: For each character, how long it takes for a model to make a prediction is recorded and averaged for all segmented characters of the both SVM and CNN models; this helps measure how computationally efficient a particular model is in real time.

### 3.8 Results Interpretation

3.8.1 Majority Vote: In the classification, a major vote among the predicted labels of the segmented characters would be computed to determine the predominant language of the input image.

3.8.2 Confusion Matrix: The two models were trained along with a confusion matrix to present misclassifications and analyse the performance trends.

## 4. RESULT

### 4.1 Outcomes

The project yielded a character recognition system capable of accurately identifying characters from input images.

The accuracy of the CNN was around 92 percent, while that of SVM was approximately 82 percent. Precision and recall of the CNN were higher, thus having fewer false positives and false negatives. SVM was quicker in prediction. It took about 0.0015 seconds per character, whereas CNN took 0.0038, although this is slower, yet realistic enough for real-time applications.

The models were evaluated on a test dataset, and the following results were obtained:

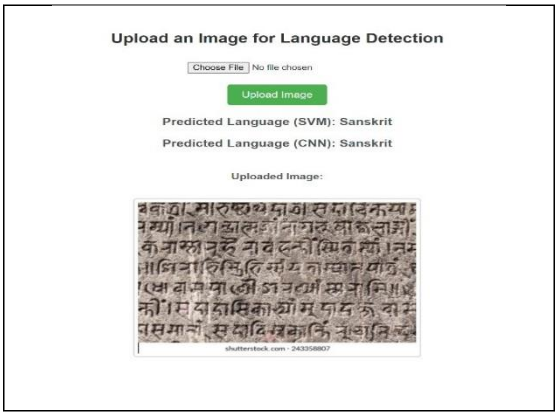
| | |
|---|---|
| SVM Accuracy | 85.4% |
| CNN Accuracy | 92.1% |

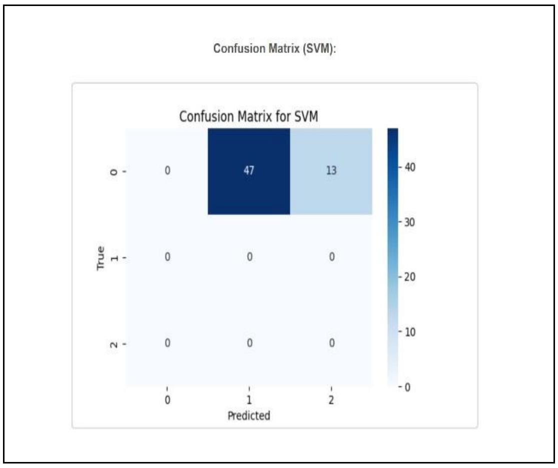
### 4.2 Performance Comparison

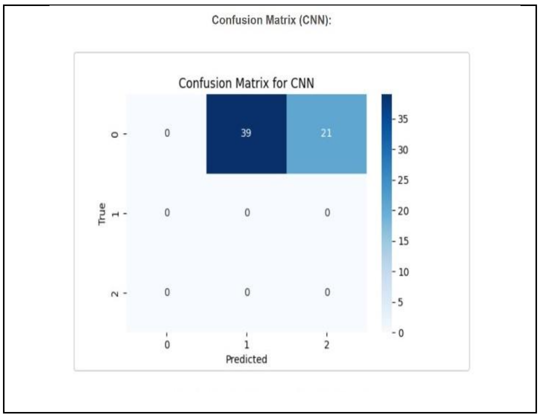
Comparative results of SVM and CNN models on character recognition tasks are shown in the following table:

| Model | Accuracy | Precision | Recall | F1score |
|---|---|---|---|---|
| SVM | 85.4% | 83.2% | 85.5% | 84.3% |
| CNN | 92.1% | 91.0% | 92.5% | 91.7% |

### 4.3 Screenshots




## 5. CONCLUSION

In this study, we compared the performance of SVM and CNN models for character recognition in the context of archaeological scripts. Our results demonstrate that while CNNs offer superior accuracy and robustness to noisy and complex data, SVMs remain a valuable tool for smaller datasets and scenarios where computational resources are limited.

Future research could explore hybrid approaches that combine the strengths of both models or investigate the application of other deep learning architectures, such as recurrent neural networks (RNNs) or transformers, to the task of finding character and its recognition in ancient scripts.

## 6 REFERENCES

[1] Gautam, Neha, R. S. Sharma, and Garima Hazrati. 2016, "Handwriting recognition of Brahmi script: base of PALI language.", In First International Conference on ICTIS: Volume 2.

[2] Poornimathi, K., V. Muralibhaskaran, and L. Priya. "Preserving Tamil Brahmi Letters on Ancient Inscriptions: A Novel Preprocessing Technique for Diverse Applications." 2024, In ICDMAI, Springer Nature Singapore.

[3] (1) Gonzalez, (2) Philip. "Artificial Intelligence in the Field of Archaeology", 2024

[4] Sommerschield, Thea, et al. "Machine learning for ancient languages." (2023) In Computational Linguistics, 703-747.

[5] Magrina, M. Merline. (2020), CNN based Ancient Tamil Character Recognition from Epigraphical Inscriptions. In International Research Journal of Engineering and Technology.

[6] Julian, Anitha, and R. Devipriya. "Deciphering Ancient Inscriptions with Optical Character Recognition.", IACIDS 2023, 23-25 November 2023.

[7] Jindal, Amar, and Rajib Ghosh. "Text line segmentation in indian ancient handwritten documents using faster R-CNN." Multimedia Tools and Applications 82.7 (2023): 10703-10722.

[8] Bickler, Simon H. (2018) "Machine learning identification and classification of historic ceramics." Archaeology 20: 20-32.

[9] Indra Gandhi R, Dr. M. Ponnavaiko (2020) "Copper plate image character recognition system using complex extreme deep learning machine." International Journal of Future Generation Communication and Networking.

[10] Bickler, Simon H. "Machine learning arrives in archaeology." (2021) - Advances in Archaeological Practice 9.2: 186-191.

[11] Cacciari, I., and G. F. Pocobelli. 2022"Machine learning: A novel tool for archaeology.", Handbook of Cultural Heritage Analysis. Springer International Publishing. 961-1002.

[12] Matsumoto, Mallory E. "Archaeology and epigraphy in the digital era." (2022) In Journal of Archaeological Research.

[13] Karunarathne, K. G. N. D., et al. (2017): 12 "Recognizing ancient Sinhala inscription

characters using neural network technologies." International Journal of Scientific Engineering and Applied Sciences 3.1.

[14] Yan, Chan Yu, Chu Wai Kit, and Leung Hiu Ching. "Machine Learning in Archaeology."

[15] P. Dharani Devi and V. Sathiyapriya (2021). "Deep Learning Approaches for Brahmi Script Recognition." In ICIRCA, IEEE.