Machine Learning-Based Approach for Classifying Damaged Characters in Early Brahmi Sri Lankan Inscriptions

G.C.J.I. Guruge[1], U.D.R.S. Thilakarathna[2], M.B. Dissanayake[1]*, A.J. Pinidiyaarachchi[3], W.I. Malsiri[4], M. Niranjan[5]

[1]*Department of Electrical and Electronic Engineering, University of Peradeniya, Peradeniya, 20400, Sri Lanka*
[2]*Department of Archaeology, University of Peradeniya, Peradeniya, 20400, Sri Lanka*
[3]*Department of Statistics and Computer Science, University of Peradeniya, Peradeniya, 20400, Sri Lanka*
[4]*Department of Archaeology, Buddhist and Pali University of Sri Lanka, Homagama, 10200, Sri Lanka*
[5]*University of Southampton, Southampton, SO17 1BJ, United Kingdom*
**maheshid@eng.pdn.ac.lk*

Early Brahmi Inscriptional character recognition for epigraphy is a computer vision task challenged by data scarcity, in which surface weathering, biological growth, and physical erosion progressively degrade the visual representation of the character. Although prior benchmarks have established performance ceilings on intact inscriptional characters, no study has quantified how those methods generalise to glyphs of varying damage severity. This work addresses that gap through a controlled external-validation protocol applied to three top-performing classical methods identified as promising in earlier work: support vector machines (SVM) with Radial Basis Function (RBF) kernel, k-nearest neighbours (k-NN), and multi-layer perceptrons (MLP). A locally collected 545-image, six-class dataset of Early Brahmi Inscriptional characters extracted from field-study was stratified into three curator-assigned damage tiers: undamaged (n=388), mildly damaged (n=61), and severely damaged (n=96), with samples from the latter two held out for testing. Histogram-of-Oriented-Gradients (HOG) features were extracted at 64×64 resolution following Contrast Limited Adaptive Histogram Equalisation (CLAHE) preprocessing. A two-stage 5-fold cross-validation protocol was applied: hyperparameters were selected by inner-loop grid search on the training pool, then evaluated through an outer 5-fold loop in which each fold's model was tested on a held-out 20% of the undamaged, mildly-damaged, and severely-damaged pools. SVM-RBF achieved the highest accuracy at every damage level (Undamaged: 0.730 ± 0.038; Mildly: 0.685 ± 0.038; Severely: 0.496 ± 0.018), with non-overlapping error bars to k-NN (Severely: 0.406 ± 0.011) and MLP (Severely: 0.408 ± 0.040). All the mildly and severely damaged image classifications were closely evaluated and validated by the domain experts. All three methods exhibited a non-linear performance cliff, averaging 17.4 percentage points between mild and severe damage, and one specific class became effectively unclassifiable across every method ($F1 < 0.10$), empirical evidence for a feature-space-level limit on classical methods that motivates the use of learned representations and image restoration as complementary approaches to closing the damage-induced accuracy gap.

Keywords: Early Brahmi Inscriptions, damaged characters, K-Fold cross-validation, classification, SVM