



CRHKS - Character Recognition in Handwritten Kadamba Script: A Dataset-Based Approach

Y.Yuktharasi, Dr. Beebi Naseeba*

ABSTRACT: The Kadamba manuscript is an early South Indian script derived from the Brahmi script developed in the fourth century by the Kadamba dynasty. It is an ancient and historically significant writing system from South India, yet its preservation and computational analysis remain challenging due to the absence of standardized digital representation. It plays a pivotal role in the development of Kannada and Telugu scripts and is frequently found in early inscriptions. This paper presents a novel dataset specifically designed for the recognition of handwritten Kadamba script. The Kadamba script data set comprises 29 consonants, 5 vowels, and 10 numerals. Data were collected from 100 participants representing a diverse range of ages and genders. Participants were provided with sample templates and instructed to write isolated characters using regular pens on A4 sheets. To emulate the appearance of traditional manuscripts, characters were written by individuals of various backgrounds. The data set collected was stored in both CSV and image formats. Each handwritten sheet was scanned and processed through a structured pipeline to enhance image quality and ensure uniformity. To preserve the structural integrity of the script, the samples underwent digitization and preprocessing steps, including adaptive binarization and contour-based segmentation. These processed samples were then used to build machine learning models. This work addresses the lack of Kadamba script resources by introducing a benchmark dataset. This contribution systematically bridges the gap by offering a standardized benchmark Kadamba script for numeral and vowel recognition, which are highly valuable in the context of manuscript analysis and research like OCR, Historical and Cultural Documentation, Restoration of Damaged Manuscripts, Semantic Analysis and Text Understanding, etc.

Key Words: Brahmi, Kadamba, binarization, OCR, semantic analysis.

Contents

1 Introduction	1
2 Historical Context	2
3 Structure of the Dataset	3
4 Data collection and Acquatisation Process	5
5 Systematic Data Processing Approach	5
5.1 Limitations and Considerations	7
6 Conclusion	7

1. Introduction

One of the earliest writing systems in use in South India is the Kadamba script. It was created during the Kadamba dynasty's rule in the fourth to sixth centuries CE [1-2]. This script, which originated from the Brahmi script, was primarily used to write early Kannada and Sanskrit. Important information such as royal orders, donations to temples, and historical events was primarily recorded using it in palm-leaf manuscripts, copper plates, and stone carvings. The Kadamba script is an ancient Indian writing system that dates back approximately 1,500 years and was used in the regions of Karnataka and Goa. Developed during the reign of the Kadamba dynasty, it was primarily inscribed on stone and copper plates to document important matters such as royal decrees and temple donations.

Notably, the development of contemporary Telugu and Kannada scripts was greatly aided by the Kadamba script. Its letters have a style that is very different from modern writing systems and are recognized for

* Corresponding author.

2010 *Mathematics Subject Classification*: 35B40, 35L70.

Submitted August 21, 2025. Published October 09, 2025

their elegant curves.

The Kadamba script had distinct regional changes as it developed structurally from the southern Brahmi script. It stands out from its northern equivalents due to its comparatively basic geometric patterns and beautifully curled letters. Many people consider this script to be the original Kannada-Telugu script, from which the current Kannada and Telugu scripts evolved. It marks a turning point in the development of writing in South India [3-4].

Nonetheless, there are a number of difficulties in studying the Kadamba script. Decipherment and interpretation are challenging due to the small number of surviving inscriptions, variances in handwriting styles, and the absence of regular orthography. Furthermore, age and environmental factors have caused numerous inscriptions to become partially eroded or destroyed. Yet there are a number of difficulties in studying the Kadamba script. Decipherable and interpretation are challenging due to the small number of surviving inscriptions, variances in handwriting styles, and the absence of regular orthography. Furthermore, age and environmental factors have caused numerous inscriptions to become partially eroded or destroyed. The extant Kadamba inscriptions, composed in Sanskrit and early Kannada, are essential for comprehending the linguistic, religious, and sociopolitical evolution of early South India, not withstanding these difficulties. They provide proof of the early usage of regional languages in governmental settings, the growth of Brahmanical Hinduism, and the support of Jainism [5-6].

Importance of the Data

- This script is notable for its influence on the development of modern Kannada and Telugu scripts. The letters are characterized by graceful curves and appear quite distinct from modern writing styles. Due to the limited number of surviving examples and variations in handwriting, deciphering the script is challenging. Inscriptions written in Sanskrit and Kannada serve as the primary sources of Kadamba history.
- There were 2900 consonant samples, 500 vowel samples, and 1000 numerical The samples in all, with each participant contributing one sample per character. This dataset, which consists of 4400 isolated character pictures, is a useful tool for creating and testing deep learning models for the recognition of handwritten characters in the Kadamba script.
- There are currently very few datasets in the Grantha script. This special dataset supports the creation of and enhancement of machine learning models for computer vision by providing one of the biggest sets of handwritten data in the Grantha script [1, 8, 9]. This dataset might be used as a reference standard by other researchers to identify Grantha numbers and vowels in handwritten characters [2-7].
- The suggested Kadamba script dataset is an essential resource for script recognition applications including machine learning and deep learning. Since the Kadamba script's historical impact on South Indian languages like Telugu, Kannada, and many more, this dataset could help with the development of language technology tools based on historical linguistic legacy in addition to research into ancient scripts.

2. Historical Context

A CapsNet-based method is used in this study to identify Devanagari characters in manuscripts. The dataset consists of 399 classes with 12 modifiers, 3 conjuncts, and 33 basic characters. Because CapsNet can record spatial relationships, it was selected. The model was evaluated across a range of epochs and train-test splits (70:10, 80:20, and 70:30). The greatest recognition accuracy of 94.6% was attained [8]. In the past to the introduction of modern numerals, this study examined how palm-leaf texts were arranged using numbers in ancient Sri Lankan monasteries. It concentrated on Sinhala numbers, which are significant to the locals and their culture. Over a period of three years, specialists assisted in the examination and interpretation of 4,568 Buddhist monastic writings. The study discovered five distinct kinds of Sinhala numerals that existed at various points in time, illustrating the evolution of the number system [9]. A dataset of 262 degraded Tamil palm leaf manuscripts from "Naladiyar", "Tholkappiyam",

and “Thirikadugam” is presented in the paper. A Nikon camera was used to take the pictures, which were then processed and binarized using Otsu thresholding. Research in ML, DL, Transfer Learning, AI, and ANN is supported by the dataset [10].

This study uses a CNN-based method to digitally recognize handwritten characters in order to preserve Tamil palm leaf manuscripts. Characters are scaled to fixed pixels for training without the need for manual feature extraction. The model outperformed conventional techniques with an accuracy of 97% on a dataset of 60 classes [11].

In order to recognize twelve Tamil vowels in palm leaf texts, this research suggests employing B-spline curve recognition. B-splines capture the different curve angles of each vowel, giving them originality and durability. The system performs better in accuracy than current methods and successfully recognizes characters from various authors. This work addresses problems including noise, punch holes, and overlapping letters by presenting a four-step process for digitizing and recognizing Tamil palm-leaf manuscripts. It is useful for real-time historical research since it can recognize 247 letters and 12 numerals with 96.04% accuracy using sophisticated preprocessing, segmentation, and a multi-class CNN for 125 classes [12].

Building strong machine learning (ML) and deep learning (DL) models for script identification and recognition requires a substantial amount of digitized Kadamba script data, which is why this dataset was compiled. Due to its unique character structure and ancient origin, the Kadamba script, which has been used historically for inscriptions and early Kannada-Telugu scripts in South India, has few digital resources.

This dataset, which offers a complete collection of handwritten Kadamba script samples, was developed to fill this gap. For ease of use in ML/DL applications, the data was carefully collected from numerous participants, digitized, and structured.

The primary contribution of this paper is a standardized dataset of handwritten Kadamba characters consisting of consonants, vowels and numerals with detailed preprocessing steps (binarization, segmentation, CSV/image formats).

While the dataset is designed to support ML/DL applications (as highlighted in Sections 4–5), our current study does not implement or evaluate any recognition models. Instead, it lays the groundwork for such future research by addressing the scarcity of digital Kadamba script resources. While we have not implemented or evaluated any recognition models in this work, we recognize the importance of such models for script analysis. The dataset we present has been carefully structured (with standardized image formats, class labels, and preprocessing) specifically to facilitate future model development by researchers in this field.

3. Structure of the Dataset

Handwritten samples of 44 different characters, consisting 29 consonants, 5 vowels, and 10 numbers, make up the Kadamba script dataset, as detailed in Tables 1 and 2,3,4,5..100 individuals, spanning a wide range of backgrounds, provided the data. On simple A4 white sheets, participants were given printed sample templates containing the Kadamba script and told to create isolated characters with ordinary pens. In order to digitize data, the handwritten papers were digitized using the cameras of handheld devices as depicted in Fig. 1.

Table 1: Kadamba Numerals and Their Representations











Numbers	0	1	2	3	4	5	6	7	8	9
Numbers sample Image										
Class Numbers	0	1	2	3	4	5	6	7	8	9

Table 2: Kadamba Vowels and Their Representation

Vowels	a	aa	i	u	e
Vowel Sample Image					
Class Numbers	10	11	12	13	14

Table 3: Kadamba Consonants and Their Representation

Consonants	ka	kha	Ga	Gha	Na	Ca	Cha	Ja	NA	Ta
Consonant Sample Image										
Class Numbers	15	16	17	18	19	20	21	22	23	24

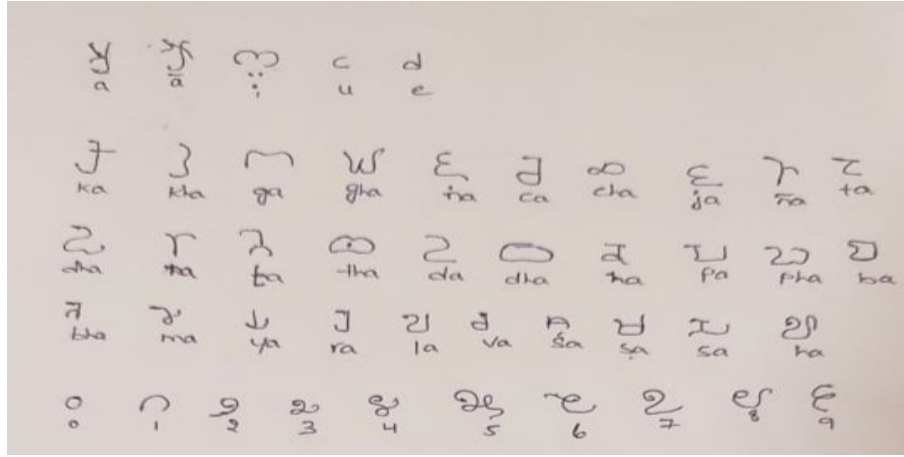


Figure 1: Standard A4 Sheet Used for Collecting Script Samples

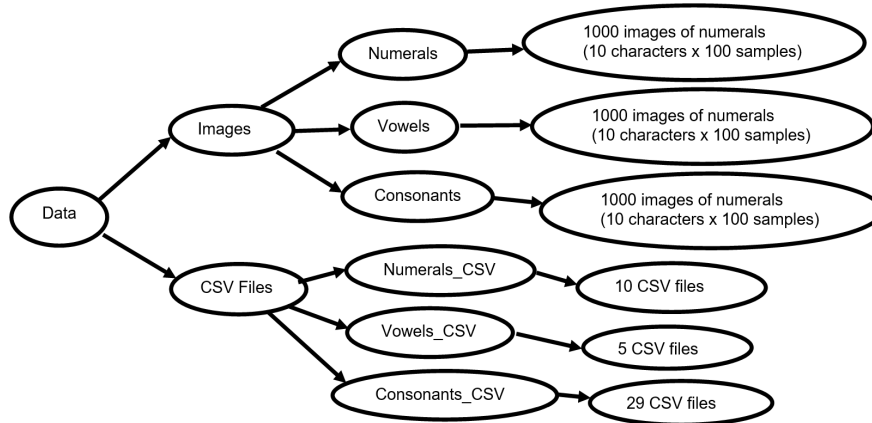


Figure 2: Kadamba Dataset Structure

Table 4: Kadamba Consonants and Their Representation




















Consonants	Dha	Na	La	Tha	Da	Dha	Na	Pa	Pha	Ba
Consonant Sample Image										
Class Numbers	25	26	27	28	29	30	31	32	33	34

Table 5: Kadamba Consonants and Their Representation

Consonants	Bha	Ma	Ya	Ra	La	Va	Sa	Sa	HA
Consonant Sample Image									
Class Numbers	35	36	37	38	39	40	41	42	43

4. Data collection and Acquatisation Process

Right after the initial gathering of data, the samples of handwritten Kadamba script were digitized using a set of techniques to create a structured digital representation. 100 participants' handwritten A4 sheets were converted to digital format using a REALME NOTE 9 PRO smartphone with a 48 MP rear camera. As a result, all character samples were captured at high resolution, enabling accurate downstream processing. After being digitized, the photos were segmented using the popular computer vision library OpenCV-Python (version 4.9.0). This automated procedure streamlined the data preparation pipeline by accurately detecting and extracting individual characters—consonants, vowels, and numbers—from the scanned sheets. "The datasets used for this study were collected from digitized inscriptions available through the Karnataka Itihasa Academy's official website". To normalize the dataset and model training, more preprocessing was done. To lower computational complexity and preserve consistency throughout the dataset, the segmented images were shrunk to uniform dimensions and turned to grayscale. Each image was examined manually to confirm the quality of the segmentation and eliminate any artifacts or irregularities. There are 4400 digital images in the collection, which includes 2900 consonants (29 characters \times 100 samples), 500 vowels (5 characters \times 100 samples), with 1000 numerals (10 characters \times 100 samples). These are properly arranged according to character class into the appropriate folders. 44 CSV files, each representing a single character class—10 of numerals, 5 for vowels, along with 29 for consonants—are also included in the collection. Fig. 2 shows the organization of these folders and files, giving a clear picture of the accessibility and data arrangement.

5. Systematic Data Processing Approach

A smartphone was used to scan all handwritten documents, which were then saved in JPG format. The bounding box technique was used to extract each of the Kadamba characters from these scanned images, enabling accurate segmentation and noise reduction. To guarantee segmentation quality and consistency, each character image was subsequently scaled to predefined dimensions of 28×28 pixels and manually verified.

Characters in the retrieved images were rendered in black on a white background after being converted to binary (black and white). The dataset was arranged into 44 different folders, each representing a single Kadamba character (consisting of 10 numerals, 5 vowels, and 29 consonants), as shown in Fig. 2. Every image was given a class label after being flattened into a vector of size 1×784 , which represents the 28×28 pixel grid. Each character has an own CSV file containing these vectors. A total of 44 CSV files were produced, comprising 10 for numerals, 5 for vowels, and 29 for consonants. A single character in each sample and its label are represented by each of the 100 rows in each file. Fig. 3 provides a visual summary of the entire data preparation pipeline, which includes digitization, segmentation, and formatting.

A systematic OCR pipeline was used to process the dataset in order to simplify the digital recognition of

handwritten Kadamba script. The purpose of this pipeline was to transform unstructured handwriting input into labeled, machine-readable samples that could be used to train deep learning models. The procedures are described below, with reference to the pipeline diagram, and are depicted in Fig. 3:

Data collection: Handwritten samples of Kadamba script characters were provided by 100 participants, representing a range of ages and genders. Ten numbers, five vowels, and 29 consonants are included in the collection.

Image Acquisition: On A4 paper sheets, participants were told to write isolated characters with regular ink pens. The dataset is based on these sheets, which represent the process of creating a traditional text.

Digital Conversion and Scanning: To maintain the script's precise structural details and clarity, high-resolution scanners were used to digitize the handwritten pages. In order to handle the analog input further, this phase transforms it into digital image files.

Image Processing: A few kinds of preprocessing techniques were used to enhance the raw data's consistency and quality:

Data cleaning: is the process of eliminating noise and scanning artifacts.

Noise reduction and enhancement: methods to highlight character qualities, such as sharpening and contrast normalization.

By selectively adjusting threshold values, adaptive binarization transforms grayscale photos into binary images while maintaining stroke details in a range of lighting environments.

Labeling: A manual annotation of the appropriate label—vowel, consonant, or numeral—was made for each segmented character. Training supervised learning models requires this labeling. For compatibility with OCR systems, annotations were kept in structured CSV files in addition to images.

Data Storage: Two formats were used to store the completed and processed dataset: JPG image format: Preserves character visual information. CSV format: Encodes labels and information known as metadata to make it suitable with machine learning.

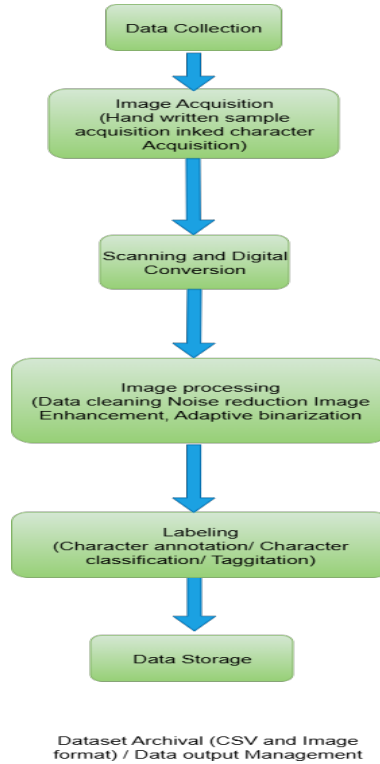


Figure 3: Character Data Pipeline for Historical Script Recognition

5.1. Limitations and Considerations

Since data collection was done by hand, handwriting styles varied, which could have an impact on recognition consistency. Furthermore, even though the dataset is vast, it might not be enough to train really sophisticated deep learning models that need more data.

6. Conclusion

This work presents a structured approach to creating a benchmark dataset of handwritten Kadamba script. It includes 30 consonants, 5 vowels, and 10 numerals, with data meticulously collected, scanned, and organized for clarity and usability. The dataset, provided in both image and CSV formats, is designed to support the training and evaluation of machine learning models for recognizing this ancient script. Future research will focus on detecting and restoring damaged or faded characters to enhance the readability of manuscripts. Efforts will also be made to integrate the dataset into digital preservation platforms for broader accessibility. Additionally, OCR models are being developed to handle noisy or incomplete scripts by combining restoration and recognition techniques.

Acknowledgments

The authors would like to express their gratitude to VIT-AP University staff and students for their invaluable assistance.

Ethical Statement:

Before starting data collection, all handwritten characters were obtained with permission from the appropriate college authorities. Since neither humans nor animals were used in the study, ethical approval was not thought to be required.

References

1. Challa, N. P., *Post Digitization Challenges and Solutions for India Palm Leaf Manuscripts*, Medicon Engineering Themes, vol. 3, no. 2, pp. 47–49, 2022. [Online]. Available: https://www.researchgate.net/publication/362430209_Post_Digitization_Challenges_and_Solutions_for_India_Palm_Leaf_Manuscripts
2. Yohoshiva, B., and Challa, N. P., *GHCR—A dataset for Grantha handwritten character recognition*, Data in Brief, vol. 56, p. 110783, 2024.
3. Dhivya, S., and Devi, U. G., *TAMIZHĪ: historical Tamil-Brahmi script recognition using CNN and MobileNet*, ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 20, no. 3, 2021.
4. Jyothi, R. L., and Rahiman, M. A., *Handwritten character recognition from ancient palm leaves using gabor based multilayer architecture: GMA*, Int. J. Appl. Eng. Res., vol. 15, no. 8, pp. 827–834, 2020.
5. Amrutha Raj, V., Jyothi, R. L., and Anilkumar, A., *Grantha script recognition from ancient palm leaves using histogram of orientation shape context*, in Proc. of 2017 Int. Conf. on Computing Methodologies and Communication (ICCMC), pp. 790–794, IEEE, 2017.
6. Jyothi, R. L., and Rahiman, A., *Comparative analysis of wavelet transforms in the recognition of ancient Grantha Script*, Int. J. Comput. Theory Eng., vol. 9, no. 4, pp. 235–241, 2017.
7. Sreeraj, M., and Idicula, M. S., *An online character recognition system to convert Grantha script to Malayalam*, Int. J. Adv. Comput. Sci. Appl., vol. 3, no. 7, 2012.
8. Moudgil, Aditi, Singh, Saravjeet, Gautam, Vinay, Rani, Shalli, and Shah, Syed Hassan, *Handwritten Devanagari manuscript characters recognition using CapsNet*, Int. J. Cogn. Comput. Eng., vol. 4, pp. 47–54, 2023.
9. Cabral, Udaya, Kumara, Lakshan Dhananyaja, and Ramanan, T., *Ancient Sinhala numeral systems discovered from palm-leaf manuscripts in Sri Lanka*, J. Univ. Librarians Assoc. Sri Lanka, vol. 28, no. 1, 2025.
10. Jailingeswari, I., and Gopinathan, S., *Tamil handwritten palm leaf manuscript dataset (THPLMD)*, Data Brief, vol. 53, p. 110100, 2024.
11. Subramani, Kavitha, and Murugavalli, S., *Recognizing ancient characters from Tamil palm leaf manuscripts using convolution-based deep learning*, Int. J. Recent Technol. Eng., vol. 8, no. 3, pp. 6873–6880, 2019.
12. Athisayamani, Suganya, Singh, A. Robert, and Athithan, T., *Recognition of ancient Tamil palm leaf vowel characters in historical documents using B-spline curve recognition*, Procedia Comput. Sci., vol. 171, pp. 2302–2309, 2020.

Y. Yuktharasi,
School of Computer Science and Engineering (SCOPE),
VIT-AP University, Amaravati, Andhra Pradesh
India.
E-mail address: yuktharasi.24phd7144@vitap.ac.in

and

Dr. Beebi Naseeba,
School of Computer Science and Engineering (SCOPE),
VIT-AP University, Amaravati, Andhra Pradesh
India.
E-mail address: beebi.naseeba@vitap.ac.in