



Faculty of Computer Science and Information Technology

**Contour-KNN Brahmi Segmentation (CKBS) and Two-Phase Enhanced
Brahmi Recognition (TREBR) Methods for Automatic Brahmi Texts
Labelling**

Neha Gautam

**Doctor of Philosophy
2024**

Contour-KNN Brahmi Segmentation (CKBS) and Two-Phase Enhanced
Brahmi Recognition (TREBR) Methods for Automatic Brahmi Texts
Labelling

Neha Gautam

A thesis submitted

In fulfillment of the requirements for the degree of Doctor of Philosophy

(Computer Science)

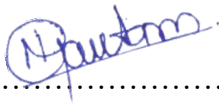
Faculty of Computer Science and Information Technology

UNIVERSITI MALAYSIA SARAWAK

2024

DECLARATION

I declare that the work in this thesis was carried out in accordance with the regulations of Universiti Malaysia Sarawak. Except where due acknowledgements have been made, the work is that of the author alone. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.



.....
Signature

Name: Neha Gautam

Matric No.: 16010146

Faculty of Computer Science and Information Technology

Universiti Malaysia Sarawak

Date: 20 Jan 2024

ACKNOWLEDGEMENT

The time spent completing my PhD has been without question one of the most challenging and rewarding experiences of my life. I have amassed a wealth of knowledge and learned much about myself. I am certain that I would not have been able to complete this journey without the support of several individuals. It is with humble gratitude that I would like to recognise them here.

First and foremost, I would like to praise and thank God, the Almighty, who has granted countless blessings, knowledge, and opportunities to the writer so that I have finally been able to accomplish the thesis.

I express my deep gratitude to my supervisor, Dr. Chai Soo See, for guiding me well throughout the research work, from title selection to finding the results. The immense knowledge, motivation, and patience of Dr. Chai have given me more power and spirit to excel in research writing. Conducting the academic study regarding such a difficult topic could not be as simple as she made this for me.

I am also indebted to the Zamalah Siswazah UNIMAS Scholarship (ZSU) for generously providing the financial support needed to pursue my study. Without their assistance, I would not have been able to progress as far as I have.

I would never forget my fellow labmates for the fun time we spent together, the sleepless nights that gave us the courage to complete tasks before deadlines, and for stimulating the discussions.

Ultimately, I am grateful to my parents, siblings, friends, and acquaintances who remembered me in their prayers for the ultimate success. I consider myself nothing without them. They gave me enough moral support, encouragement, and motivation to accomplish my goals. This thesis is dedicated to my father.

ABSTRACT

Automatic word recognition problem can be solved using an optical character recognition (OCR) system. Few studies have been seen in the field of Brahmi word recognition especially identifying compound characters and words with good accuracy. However, existing Brahmi text recognition studies have primarily relied on local datasets, hampering the standardization of datasets. To address this, the study proposes a systematic dataset creation process that encompasses data collection, pre-processing, segmentation, data augmentation, recognition, storage, and labelling. The process is initiated with data collection from diverse online sources, yielding 217 text and word samples and 801 isolated characters and compound characters. However, these samples lack uniformity in text and word sizes. The subsequent phase focuses on character isolation from words and text, utilizing a novel segmentation approach as a crucial precursor to system training. A Contour-KNN Brahmi Segmentation (CKBS) for character and compound character segmentation is introduced. Object detection identifies characters, including dots (.), and links them to their nearest left character using KNN. This approach greatly enhances segmentation, achieving an impressive 98.19% average accuracy. The segmentation approach generates 40 samples per class across 170 classes, with a 75:25 training-testing split (30 and 10 samples for training and testing, respectively). Furthermore, data augmentation techniques, including adjustments, deformations, blurring, translations, and noise introduction, are applied to enhance dataset quality and quantity. Data augmentation results in 180 training and 60 testing samples per class, improving both size and quality. Subsequently, a Two-Phase Enhanced Brahmi Recognition (TPEBR) is employed, distinguishing between global and local feature recognition. Various deep learning architectures are evaluated for classification, with resizing to meet specific input size requirements. SqueezeNet emerges as the most

effective, achieving a minimal 0.237 loss and an exceptional 97.58% accuracy. It excels in precision, recall, and F1-score. In contrast, ResNeXt Small underperforms with higher loss and lower accuracy. Comparing the Two-Phase Enhanced Brahmi Recognition (TPEBR) to the existing approaches, the Two-Phase Enhanced Brahmi Recognition (TPEBR) achieves 97.58% accuracy, while the existing approaches records 80.20% and 90.24%. Recognized characters are then organized into folders according to their recognized class, and done labelling by using Brahmi Unicode, although this step does not impact performance of the system.

Keywords: Brahmi script, deep learning architecture, data augmentation, KNN, object detection, two-phase approach, word recognition

Segmentasi Brahmi KNN Kontur (CKBS) dan Kaedah Pengiktirafan Brahmi Dua Fasa yang Dipertingkatkan (TREBR) untuk Pelabelan Teks Brahmi Automatik

ABSTRAK

Masalah pengenalan perkataan secara automatik boleh diselesaikan dengan menggunakan sistem pengenalan aksara optik (OCR). Beberapa kajian telah dilihat dalam bidang pengenalan perkataan Brahmi terutamanya mengenal pasti aksara gabungan dan perkataan dengan ketepatan yang baik. Walau bagaimanapun, kajian pengenalan teks Brahmi yang sedia ada kebanyakannya bergantung pada set data tempatan, menghalang piawaian set data. Untuk menangani ini, kajian ini mencadangkan proses penciptaan set data yang sistematik yang merangkumi pengumpulan data, pra-pemprosesan, segmentasi, penambahbaikan data, pengenalan, penyimpanan, dan pelabelan. Proses ini dimulakan dengan pengumpulan data dari pelbagai sumber dalam talian, menghasilkan 217 sampel teks dan perkataan dan 801 aksara terasing dan aksara gabungan. Namun, sampel ini kekurangan keseragaman dalam saiz teks dan perkataan. Fasa seterusnya memberi tumpuan kepada pengasingan aksara dari perkataan dan teks, menggunakan pendekatan segmentasi baru sebagai pendahuluan penting kepada latihan sistem. Satu Contour-KNN Brahmi Segmentation (CKBS) untuk segmentasi aksara dan aksara gabungan telah diperkenalkan. Pengesanan objek mengenal pasti aksara, termasuk titik (.), dan mengaitkannya dengan aksara kiri terdekat menggunakan KNN. Pendekatan ini meningkatkan segmentasi dengan ketara, mencapai ketepatan purata 98.19%. Pendekatan segmentasi menghasilkan 40 sampel setiap kelas di 170 kelas, dengan bahagian latihan-ujian 75:25 (30 dan 10 sampel untuk latihan dan ujian, masing-masing). Selain itu, teknik penambahbaikan data, termasuk penyesuaian, deformasi, kabur, terjemahan, dan pengenalan bising, diterapkan untuk meningkatkan kualiti dan kuantiti set data. Penambahbaikan data menghasilkan 180 sampel

latihan dan 60 sampel ujian setiap kelas, meningkatkan saiz dan kualiti. Kemudian, kaedah *Two-Phase Enhanced Brahmi Recognition (TPEBR)* yang diperkenalkan dalam kajian ini telah digunakan untuk membezakan antara pengenalan ciri global dan tempatan. Pelbagai seni bina pembelajaran mendalam dinilai untuk pengelasan, dengan mengubah saiz untuk memenuhi keperluan saiz input tertentu. *SqueezeNet* muncul sebagai yang paling berkesan, mencapai kerugian minimum 0.237 dan ketepatan luar biasa 97.58%. Ia cemerlang dalam ketepatan, ingatan, dan skor *F1*. Sebagai perbandingan, *ResNeXt Small* kurang berprestasi dengan kerugian yang lebih tinggi dan ketepatan yang lebih rendah. Membandingkan *Two-Phase Enhanced Brahmi Recognition (TPEBR)* dengan pendekatan yang sedia ada, *Two-Phase Enhanced Brahmi Recognition (TPEBR)* mencapai ketepatan 97.58%, manakala pendekatan yang sedia ada mencatat 80.20% dan 90.24%. Aksara yang dikenali kemudian dianjurkan ke dalam folder mengikut kelas yang dikenali, dan dilakukan pelabelan dengan menggunakan *Unicode Brahmi*, walaupun langkah ini tidak memberi kesan kepada prestasi sistem.

Kata kunci: Skrip brahmi, seni bina pembelajaran mendalam, penambahbaikan data, KNN, pengesanan objek, pendekatan dua fasa, pengenalan perkataan.

TABLE OF CONTENTS

	Page
DECLARATION	i
ACKNOWLEDGEMENT	ii
ABSTRACT	iii
<i>ABSTRAK</i>	v
TABLE OF CONTENTS	vii
LIST OF TABLES	xi
LIST OF FIGURES	xiii
CHAPTER 1 INTRODUCTION	1
1.1 Background Study	1
1.1.1. Online Text Recognition	3
1.1.2. Offline Text Recognition	3
1.2 Problem Statement	6
1.3 Research Questions	8
1.4 Objectives	9
1.5 Hypothesis	10
1.6 Scope	11
1.7 Research Methodology	12
1.8 Organization of the Thesis	13

CHAPTER 2 WRITING SYSTEMS AND SCRIPTS OF THE WORLD	16
2.1 The Writing System of the World	16
2.1.1 Logographic System	16
2.1.2 Syllabic System	16
2.1.3 Abjad System	16
2.1.4 Abugida System	17
2.1.5 True Alphabetic	18
2.1.6 Featural System	18
2.2 Overview of the Brahmi Script	20
2.2.1 Properties of Brahmi Script	20
2.2.2 Feature of Brahmi Words	22
2.3 Summary	25
CHAPTER 3 LITERATURE REVIEW	26
3.1 Brahmi Characters and Words Recognition System	26
3.2 Dataset	28
3.3 Segmentation	32
3.4 Approaches for Brahmi Text Recognition: An Overview	37
3.4.1 Holistic and Analytic Approaches	37
3.4.2 Single-phase and two-phase approach	40
3.5 Summary	45

CHAPTER 4 METHODOLOGY	46
4.1 Data Collection	46
4.2 Pre-processing	48
4.2.1 Binarization/Thresholding	48
4.2.2 Noise Removal	51
4.3 Contour-KNN Brahmi Segmentation (CKBS) Method	53
4.3.1 Object Detection	53
4.3.2 Steps of KNN	58
4.4 Data Augmentation	60
4.5 Two-Phase Enhanced Brahmi Recognition (TPEBR) Method	64
4.6 Data Storage and Labelling	70
4.7 Summary	73
CHAPTER 5 EXPERIMENTAL RESULTS	74
5.1 Result Contour-KNN Brahmi Segmentation (CKBS) (First Part)	76
5.2 Data Augmentation	78
5.3 Result of Two-Phase Enhanced Brahmi Recognition (TPEBR) Method (Second part)	78
5.3.1 Performance Metrics	79
5.3.2 Result of the Recognition Model	80
5.3.3 Comparison the performance of Two-Phase Enhanced Brahmi Recognition (TPEBR) Method with Previous Studies	90

5.4	Overall Performance of Proposed System	91
5.5	Statistical Distribution	91
5.6	Summary	92
CHAPTER 6 CONCLUSION AND RECOMMENDATION		94
6.1	Contributions	95
6.2	Limitations	97
6.2.1	Dataset	98
6.2.2	Segmentation	98
6.2.3	Brahmi Text Recognition System	98
6.3	Future Works	98
6.3.1	Dataset	98
6.3.2	Segmentation	99
6.3.3	Brahmi Text Recognition System	99
REFERENCES		100
APPENDICES		115

LIST OF TABLES

	Page
Table 2.1 Example of Compound Characters	24
Table 3.1 Comparison of the Three Approaches in Terms of Brahmi Script	39
Table 4.1 Summary of the Threshold Value is Determined Through Otsu's Thresholding Method for all of the Images	50
Table 4.2 Analysis of data augmentation techniques	61
Table 4.3 Statistics Value of Q_{11} , Q_{21} , Q_{12} , and Q_{22}	69
Table 4.4 Standard Size of each Architecture	70
Table 5.1 Performance of TPEBR Approach where Acc, Pre, Rec, F1 Represents Accuracy, Precision, Recall and F1-Score	81
Table 5.2 Performance of TPEBR Approach on Adjusted Data Augmentation Testing Sample Where Acc Represents the Accuracy	82
Table 5.3 Performance of TPEBR Approach on Blurred Data Augmentation Testing Sample where Acc Represents the Accuracy	83
Table 5.4 Performance of TPEBR Approach Approach on Deformed Data augmentation Testing Sample	84
Table 5.5 Performance of TPEBR Approach on Noisy Data Augmentation Testing Sample	85
Table 5.6 Performance of TPEBR Approach on Translated Data Augmentation Testing Sample	86
Table 5.7 Performance of TPEBR Approach on Original Data Testing Sample	87
Table 5.8 Comparison of Training and Testing Accuracy of TPEBR Approach where Acc represnts Accuracy	88
Table 5.9 The performance TPEBR Approaches of the Testing Dataset of Segmentation Approach	89

Table 5.10 Comparison of Existing and Proposed Brahmi Word Recognition Systems

90

LIST OF FIGURES

	Page	
Figure 1.1	Flow Diagram of the Brahmi Text Recognition System	6
Figure 1.2	Connection between Problem Statements and Research Questions	9
Figure 1.3	Connection between Research Questions and Research Objectives	10
Figure 1.4	Connection between Research Objective and Hypothesis	11
Figure 2.1	The Writing Systems of World (Ghosh et al., 2010; Obaidullah et al., 2018)	19
Figure 2.2	Characters and Vowels of the Brahmi Script (Roy & Mandal, 2016)	20
Figure 2.3	Vowels of the Brahmi Script (Roy & Mandal, 2016)	21
Figure 2.4	Example of Consonant and Compound Characters (Roy & Mandal, 2016)	21
Figure 2.5	Example Text in Brahmi Script (Rajan, 2010)	21
Figure 2.6	Example of Complex Compound Characters	22
Figure 4.1	A Diagram of the Overall Methodology	47
Figure 4.2	Samples of Brahmi Text, With and Without Noise	48
Figure 4.3	Input image	51
Figure 4.4	Output image of Erosion structuring element	52
Figure 4.5	Output image of Dilation structuring element	52
Figure 4.6	Input image	56
Figure 4.7	Output of object detection part	56
Figure 4.8	After merging the dot character with the relevant character	58
Figure 4.9	Example of all segmented character	59
Figure 4.10	Hierarchy of the Data Augmentation	60
Figure 4.11	Consonant, Vowels, and Compound Characters of the Brahmi Script	65

Figure 4.12	A Schematic Representation of Proposed Approach Two-Phase Enhanced Brahmi Recognition (TPEBR)	66
Figure 4.13	Process of Proposed Approach (Two-Phase Enhanced Brahmi Recognition (TPEBR)) with an Example	67
Figure 4.14	All Isolate Characters are Stored into Groups by using Segmentation and Recognition Step	70
Figure 4.15	Labeling of few Samples of Brahmi Dataset	71
Figure 4.16	Flow of Methodology of Proposed System to do Automatic Labelling along with Training and Testing Details	72
Figure 5.1	Flow of the Performance Evaluation of the Proposed System	75

CHAPTER 1

INTRODUCTION

This section provides an outline to develop an automated system to generate a Brahmi dataset, with a specific focus on the importance of automated word labelling in the context of the Brahmi script. It delves into the exploration of an appropriate methodology for future investigations. Furthermore, it addresses the formulation of problem statements and the establishment of research objectives. The structure of this chapter encompasses five key segments. The initial segment (1.1) furnishes the historical backdrop of the study, while the subsequent portions (1.2, 1.3, 1.4, and 1.5) elucidate the problem statements, scope of the study, research questions, and objectives, respectively. Additionally, the development of a hypothesis is presented in Section 1.6. The chapter concludes by delving into the research methodology in Section 1.7 and providing an overview of the thesis organization in Section 1.8.

1.1 Background Study

The modern world is becoming more digitised with the prevalence of computers, PDA screens, and devices, which are traditional books and newspapers. Carrying around large amounts of paper will require more maintenance because the content of the paper can disappear with time and under different weather conditions, making it tough to restore the removed content. Therefore, books, newspapers, or any sort of paper are now being converted into digital format via a scanning process to store them for a long time. However, scanned images are not automatically readable and editable by computers, so the text in the image cannot be used directly for further work. This issue can be solved using suitable recognition systems, which are capable of reading the content similar to the human ability

to read. Up until now, various types of recognition systems have been developed for automatic word recognition.

Numerous works in the field of English and Roman script recognition have been done compared to South and Central Asian scripts because South and Central Asian scripts are more cursive compared to English and Roman scripts (Gautam et al., 2016; Roy et al., 2016). Ancient script recognition of South and Central Asian scripts has brought a new challenge because of the significant variations in the structure of the characters, lack of resources, controversy of statements, etc (Sulaiman et al., 2019).

The Brahmi script was an ancient script that was once used in South Asia, Southeast Asia, and East Asia. A study identified 198 different contemporary scripts, such as Devanagari, Bangla, Gurmukhi, Gujarati, Oriya, etc., all descended from the Brahmi script (Gautam & Chai, 2017; Acharya, 2018). Hence, it can be said that the Brahmi script has a rich background. The Brahmi script has its origins in Theravada Buddhism, which is known as the “pillars of Asoka”, where the text on the pillars is written in Brahmi (Gautam & Chai, 2017). Similarly, text related to the Buddhist culture is also written using the Brahmi script, mainly appearing on various pillars, caves, and stones. These pillars, caves, and stones were the main ways of spreading Buddhist theology all over Asia, mostly in South Asia, Southeast Asia, and East Asia (Igunma, 2018).

Some studies in the realm of Brahmi text recognition have delivered a good performance. These investigations initially focused on the extraction of pertinent features from Brahmi text. Subsequently, the extracted features were subjected to classification via suitable algorithms. Diverse methods for feature extraction and classification were employed. For instance, Geometric feature extraction (Gautam & Chai, 2017) and zoning techniques (Gautam et al., 2016; Vellingiriraj et al., 2016) were utilized to acquire features

of Brahmi characters and Brahmi text. In contrast, a combination of rule-based systems (Gautam & Chai, 2017), template matching (Gautam et al., 2016), and Artificial Neural Networks (ANN) (Vellingiriraj et al., 2016) served as classifiers to categorize the extracted features.

The challenge of Brahmi text recognition falls within the broader domain of automatic word recognition, which can be primarily categorized as online and offline text recognition. When data is captured from paper via optical scanners or cameras, this process is known as offline text recognition. Meanwhile, a system that uses digitisers to capture written content at the time of writing is called online text recognition.

1.1.1 Online Text Recognition

In this process, words are recognised in real-time as soon as the content is written, and thus, it incorporates temporal information. Online systems have replaced the role of the pen as a function of time directly from the interface. The online recognition system is usually done through pen-based interfaces where the writer writes with a special pen on an electronic tablet.

1.1.2 Offline Text Recognition

Offline text recognition systems follow a particular process. Firstly, the text must be written on paper. The paper is then scanned with a scanner or a camera. Following the capture and scanning of the text in picture format, optical character recognition (OCR) is used to transform the image into a readable and editable format (Das et al., 2019; Sharma & Mudgal, 2019). OCR is a technology that allows for the conversion of a variety of documents, such as scanned paper documents and PDF files or pictures, into editable and searchable data. (Chaudhuri et al., 2017; Budhi et al., 2018).

The Brahmi script is an old script with a lot of written information. In the world of today, this script is no longer used for any sort of communication. Hence, offline text recognition is the most useful solution for recognising Brahmi script, with OCR employed to locate Brahmi text. With offline word recognition, analytical and holistic approaches are used to develop the Brahmi text recognition system.

1.1.2.1 Analytical Process

In the analytical process, a word is initially segmented into sub-units called characters, and then each sub-unit is categorised into classes. This step helps to train and test the recognition system. Most word recognition solutions were developed based on the analytical process.

1.1.2.2 Holistic Process

The holistic approach considers a word as a single and unified unit. Thus, the text is only segmented into meaningful words, and all words are categorised into classes. This step will help to build the system. This approach addresses the problems of the fixed or limited lexicon. Besides that, this approach can be used to reduce the lexicon in large vocabulary problems (Madhvanath et al., 1999; Madhvanath & Govindaraju, 2001).

In the comparison of both approaches, the analytical process is useful with ancient scripts where not a lot of text is available of the respective script. However, the major problem with the analytical process is that all characters are separately recognised, so word prediction might not be done because word predictions are only possible if the whole word has been recognised. Word prediction is used to increase the accuracy of the recognition system.

The holistic approach can perform well in popular scripts such as English, Arabic, and Chinese as these are the modern scripts and many dictionaries are available which can help to do word prediction.

For Brahmi text recognition, available samples of Brahmi words and text are limited, so the holistic approach might not be more suitable than the analytical approach for a Brahmi text recognition system. The holistic system also requires the use of a dictionary or book that can help to provide information about all possible words or vocabularies. However, from the review conducted in this study, no official dictionary for Brahmi words and no such book that could help to collect information regarding the meaning of the words in the Brahmi script could be found. Therefore, OCR based on the analytical process is more suitable for building the Brahmi text recognition system in this study.

Input, pre-processing, segmentation, feature extraction, and classification are the main steps employed by the OCR system. Pre-processing enhances the quality of the input image. The segmentation step isolates the characters from the text and word and is optional. In scenarios where individual characters are already separated, the segmentation process becomes unnecessary. However, for systems focusing on text recognition, this step gains significance. It follows a sequence where segmentation starts from the line, then progresses to the word level, and finally culminates with the segmentation of individual characters. Furthermore, feature extraction and classification are the decision-making steps of the OCR. Two types of paradigms can be followed to complete the feature extraction and classification steps for offline Brahmi words: supervised and unsupervised. Supervised learning identifies an unknown pattern as a member of the predefined category, while unsupervised learning groups input patterns into several clusters - hereafter defined as classes. Various methods are used to extract the features of each class, and some classifiers are also used to classify

the extracted features in the supervised learning method. In contrast, in supervised learning, there are no explicit target outputs or environmental evaluations associated with each input; rather, previous biases about which features of the input structure should be represented in the output are brought to bear in unsupervised learning.

The methods for developing a Brahmi text recognition system are illustrated in Figure 1.1.

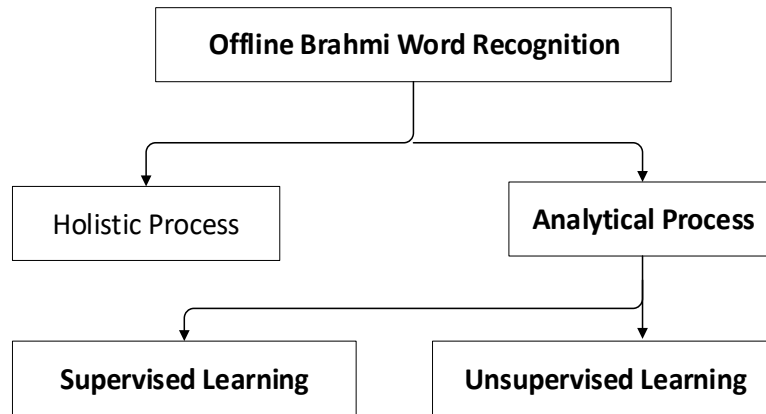


Figure 1.1: Flow Diagram of the Brahmi Text Recognition System

1.2 Problem Statement

The dataset of a study has always played an important role in word recognition systems, as it helps to train the model and if the dataset is publicly available, it provides a transparent platform for comparing the results of one study to that of other studies. In 2016, Vellingiriraj et al. (2016), who worked with Brahmi text recognition, also noticed that an official dataset would be required for Brahmi script recognition to evaluate the performance of the Brahmi script recognition algorithms. Currently, researchers investigating Brahmi script recognition have used local datasets with limited samples, which are unsuitable for further studies as such datasets are not available in the public domain. Accordingly, a different Brahmi script dataset was used in each study. A publicly available dataset should

be used as a transparent platform for a more reliable comparison of the different recognition algorithms. Thus, such studies require a Brahmi word dataset.

Furthermore, as previously emphasized, an analytical approach stands as a more viable choice for Brahmi text recognition, primarily due to the significance of a comprehensive dictionary. Consequently, the dataset must meticulously organize segmented characters within dedicated directories. Such a dataset holds the potential for advancing recognition tasks in subsequent stages. Nevertheless, the Brahmi script being an ancient writing system, no longer serves as a means of communication in the contemporary era. Therefore, automating both tasks could offer researchers valuable support in the times ahead. For the automated labelling of individual characters, the initial step involves character recognition. This signifies that the complete workflow, spanning from data collection and organized storage in the relevant directory, commences with pre-processing, segmentation, recognition of Brahmi words, and culminates in the process of automatic labelling.

Numerous segmentation methods have been developed for scripts, such as English (Kaur & Singh, 2016), Arabic (Ghaleb et al., 2017), and Devanagari (Jindal & Kumar, 2018), which aid in the development of recognition systems. Similarly, few techniques have been proposed for isolating Brahmi characters from Brahmi text. However, these studies have not been successful in segmenting all types of characters. For instance, Gautam et al. (2016) were unable to isolate characters with dot (.) features. Additionally, Singh and Kushwaha (2019) recently introduced a Brahmi text segmentation algorithm, but it did not achieve satisfactory performance due to the unique features of the Brahmi script. As a result, this remains an open research problem (Singh & Kushwaha, 2019).

Furthermore, an automated Brahmi word recognition system can be helpful for understanding each word of the Brahmi text. Machine learning can be useful to achieve such

an automated system (Dargan et al., 2020). Existing works that have developed Brahmi script recognition systems and other recognition systems should serve as a foundation for new investigations into the recognition of Brahmi script. Various researchers such as Siromoney et al. (1983), Soumya and Kumar (2015), Vellingiriraj et al. (2016) have been working to recognize the Brahmi word. However, Vellingiriraj et al. (2016) mentioned that the performance of the existing Brahmi word recognition may be further improved.

Moreover, as previously emphasized, the Brahmi script is understood by very few researchers. Therefore, automating the labelling of the dataset can be a valuable step to assist future researchers in improving both the quality and quantity of the dataset. Hence, after recognizing Brahmi characters and compound characters, automatically organizing them into a dictionary and applying automatic labels to each entry becomes an essential task that has not been addressed in the context of the Brahmi script.

1.3 Research Questions

Based on the issues presented in the problem statement, several research inquiries emerge, with a specific focus on aspects such as the dataset, segmentation, and recognition procedures. These research questions are denoted by the abbreviation "RQ" signifying their prominence in the study.

RQ1: How can a dataset be effectively developed for the Brahmi text recognition system?

RQ2: What automated procedures can be implemented for labelling and storing data within the Brahmi text recognition dataset?

RQ3: What approach can be devised to efficiently segment various character categories present within the Brahmi text?