



# **AI-Powered Sinhala Character Recognition and Digital Transformation**

R. A. Sumudu Vidyalkara  
Reg. No.: MS12906468

A THESIS  
SUBMITTED TO  
SRI LANKA INSTITUTE OF INFORMATION TECHNOLOGY  
IN PARTIAL FULFILMENT OF THE REQUIREMENTS  
FOR THE DEGREE OF  
MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

December 2024

I certify that I have read this thesis and that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

.....  
Prof. Anuradha Jayakody (Supervisor)

Approved for MSc. Research Project:

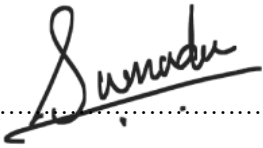
.....  
MSc. Programme Co-ordinator, SLIIT

Approved for MSc:

.....  
Head of Graduate Studies, FoC, SLIIT

# DECLARATION

This is to certify that the work is entirely my own and not of any other person, unless explicitly acknowledged (including citation of published and unpublished sources). The work has not previously been submitted in any form to the Sri Lanka Institute of Information Technology or to any other institution for assessment for any other purpose.

Sign: .....  
  
Sumudu Vidyalkara

2025-01-07  
Date: .....

# ABSTRACT

## AI-Powered Sinhala Character Recognition and Digital Transformation

Sumudu Vidyalkara

MSc. in Information Technology

**Supervisor:** Prof. Anuradha Jayakody

December 2024

This research focuses on developing an effective system for recognizing and converting handwritten and printed Sinhala text into digital format. As the primary language of Sri Lanka, Sinhala presents unique challenges for handwriting recognition due to its intricate strokes and complex character structures. Existing methods often fall short in accurately interpreting Sinhala characters, highlighting the need for a tailored solution. The proposed system employs Convolutional Neural Networks to classify and recognize Sinhala characters with high precision. A key innovation is error-guided preprocessing, applied iteratively to images misclassified during the initial training phase. Failed images are processed using methods such as blurriness detection, dynamic contrast adjustment, noise removal with bilateral filtering, and morphological operations for stroke enhancement. This approach ensures improved image quality and meaningful feature extraction for subsequent retraining. Additional techniques like contour analysis and gradient-based feature extraction further enhance the system's recognition capabilities. To optimize performance, strategies such as data augmentation, hyperparameter tuning, and model ensembles are explored, improving the system's adaptability and robustness. The system is evaluated on a diverse dataset of handwritten and printed Sinhala text, demonstrating significant improvements in recognition, accuracy and efficiency. Its applications include optical character recognition, document digitization, and automated form processing. This thesis contributes a comprehensive, CNN-based methodology tailored to the complexities of Sinhala script, offering a promising solution for advancing Sinhala language technologies.

# **ACKNOWLEDGEMENT**

I would like to express my profound appreciation to Professor Anuradha Jayakody for his great mentoring, steadfast support, and helpful counsel during the whole process of completing my thesis. His profound knowledge and mastery in the topic have had a substantial impact on my study, greatly enhancing its quality. I am really grateful for his support and commitment, which have played a crucial role in assisting me in understanding the intricacies of scholarly writing.

I would like to sincerely thank the academic members of SLIIT for their profound knowledge and valuable contributions that have significantly influenced the progress of this thesis.

Finally, I am grateful to my family and friends for their consistent support, comprehension, and motivation during my academic pursuit. Their unwavering faith in me has served as a wellspring of resilience and drive.

This thesis is the result of the assistance and direction given by Professor Anuradha Jayakody and others, and I am genuinely appreciative of it.

# TABLE OF CONTENTS

DECLARATION .....	ii
ABSTRACT .....	iii
ACKNOWLEDGEMENT .....	iv
TABLE OF CONTENTS .....	v
List of Figures .....	<b>Error! Bookmark not defined.</b>
List of Tables .....	viii
Chapter 1 - Introduction .....	1
1.1 Background .....	1
1.2 Problem Statement .....	2
1.3 Research Questions & Objectives .....	3
1.3.1 How do the strengths, weaknesses, and limitations of current methods impact Sinhala Character recognition? .....	3
1.3.2 How can Convolutional Neural Networks (CNNs) be structured and what preprocessing techniques can be applied to effectively recognize and handle the intricate features and complexities of Sinhala characters? .....	4
1.3.3 How can a prototype system be implemented to effectively process both handwritten and printed Sinhala characters, and what data preparation and augmentation optimizations can improve its performance? .....	5
1.3.4 How can the proposed system be evaluated for accuracy and efficiency in recognizing Sinhala handwriting, and how does it perform across varied handwriting styles, qualities, and real-world scenarios? .....	6
1.4 Significance of the Study .....	6
1.5 Overview .....	7
Chapter 2 - Literature Review .....	9
2.1 Artificial Intelligence in Handwriting Recognition .....	9
2.2 Limitations and Challenges .....	16
2.2.1 Limited Availability of High-Quality Annotated Datasets .....	16
2.2.2 Variability in Handwriting styles and quality .....	16
2.2.3 Computational Requirements .....	17
2.2.4 Real-Time Recognition .....	17
2.2.5 Privacy & Security .....	18
Chapter 3 - Methodology .....	19
3.1 Data Collection and Annotation .....	19
3.1.1 Objective .....	20
3.1.2 Dataset Sources .....	21
3.1.3 Annotation Methodology .....	21
3.1.4 Dataset Diversity .....	22
3.2 Data Preprocessing .....	22
3.3 Model Design and Development .....	23
3.3.1 Convolutional Layers and Filters .....	24

3.3.2 Kernel Size .....	24
3.3.3 Pooling Layers.....	24
3.3.4 Dropout Layers.....	24
3.3.5 Dense Layer and Neurons .....	24
3.3.6 Activation Functions .....	24
3.3.7 Optimizer and Learning Rate.....	25
3.4 Initial System Performance Evaluation .....	25
3.4.1 Dataset Utilization .....	25
3.4.2 Evaluation Metrics .....	25
3.5 Error-Guided Preprocessing and Iterative Refinement.....	26
3.5.1 Blurriness Detection: .....	26
3.5.2 Dynamic Contrast Limited Adaptive Histogram Equalization (CLAHE) .....	26
3.5.3 Noise Removal .....	27
3.5.4 Stroke Enhancement.....	27
3.6 Prototype System Development and Integration .....	27
Chapter 4 - Implementation: System Development and Integration.....	28
4.1 Data Collection, Initial-Preprocessing and Annotation .....	29
4.1.1 Data Collection.....	29
4.1.2 Initial-Preprocessing .....	29
4.1.3 Data Annotation.....	31
4.2 Initial Model Training and Validation.....	32
4.2.1 Model Architecture.....	32
4.2.2 Training .....	34
4.2.3 Model Validation.....	35
4.3 Error-Guided Image Preprocessing.....	37
4.4 Retraining the Model .....	43
Chapter 5 - Testing, Evaluation & Validation .....	44
5.1 Performance of the Training Set.....	45
5.2 Performance of the Validation Set.....	46
5.3 Performance of the Test Set .....	46
5.4 Analysis of the Retraining Procedure .....	47
5.5 Evaluation against Initial Model.....	47
5.6 Validation against Existing AI Systems .....	48
5.6.1 Test Case 1 .....	48
5.6.2 Test Case 2 .....	49
5.7 Model Integration.....	51
5.7.1 Integration Test 1.....	51
5.7.2 Integration Test 2.....	52
5.7.3 Detailed Observations.....	54
5.7.4 Key Highlights .....	55
5.8 Summary .....	56
Chapter 6 - Conclusion.....	58
6.1 Research Questions Addressed .....	58

6.1.1	How can a deep learning technique like CNN be used to accurately and efficiently recognize handwritten Sinhala text? .....	58
6.1.2	How can challenges such as limited annotated datasets and handwriting variability be overcome? .....	58
6.1.3	Can a system be developed to handle both real-time and offline scenarios? .....	59
6.2	Objectives Assessed.....	59
6.3	Key Contributions and Impact .....	60
6.4	Limitations and Future Work .....	60
6.5	Summary .....	61
References .....		62
APPENDIX .....		64
Appendix 1: Validation Results.....		64



# List of Tables

Table 1 - Results Summary .....	45
---------------------------------	----