An Efficient Approach For The Translation Of Brahmi Script Using OCR

1st Khushi Singh
Computer Science and Engineering
Graphic Era Hill University
Dehradun, India
khushisingh5716@gmail.com

2nd Yash Kashyap
Computer Science and Engineering
Graphic Era Hill University
Dehradun, India
yashkashyap1110@gmail.com

3rd Shagun Semwal
Computer Science and Engineering
Graphic Era Hill University
Dehradun, India
shagunsemwal3 @gmail.com

4th Vishal Ansari
Computer Science and Engineering
Graphic Era Hill University
Dehradun, India
vishal.ansari998877 @gehu.ac.in

5th Dr.Satvik Vats
Computer Science and Engineering
Graphic Era Hill University
Dehradun, India
?@gmail.com

***Abstract*—This work aims to overcome linguistic and historical barriers by translating the ancient Brahmi script into modern languages using Optical Character Recognition (OCR) technology. Because of its complex characters and varied historical forms, the Brahmi script, which has great historical and cultural significance, presents particular difficulties in transcription and translation. Acknowledging the necessity of accessing the rich cultural legacy and amount of information contained in this script, our research focuses on creating and utilizing an OCR model that is specifically suited to the subtleties of Brahmi script recognition. The principal aim of this study is to assess the model's ability to accurately translate Brahmi script into a target language after transcription. We have assembled a large dataset of Brahmi script examples from a variety of historical backgrounds and linguistic situations as part of our technique. The characters in the script are then decoded using the OCR model, which was specially created for Brahmi script recognition. The identified Brahmi text is then translated into the target modern language using a translation algorithm. Our study intends to demonstrate the precision and effectiveness of OCR technology in translating Brahmi script through a thorough review, highlighting its potential to protect cultural heritage and promote linguistic accessibility. This project has potential uses in digital preservation, cross-cultural communication, and education, in addition to being beneficial for linguists and historians.**

***Index Terms*—OCR, Brahmi Script, Translation.**

I. Introduction

One important historical and cultural relic is the Brahmi script, an old writing system that dates back to the Indian subcontinent in the sixth century BCE. Its complex characteristics and range of historical forms make correct transcription and subsequent translation extremely difficult, necessitating advanced technological solutions. In response, the goal of this research is to leverage the revolutionary potential of cutting-edge optical character recognition (OCR) technology to further the discipline. Our objective is to decipher the intricacies of the Brahmi script and translate it into modern languages with ease, so making a significant contribution to linguistic and historical research as well as providing insightful information for digital preservation, intercultural dialogue, and educational endeavours.

A. The Need of Translation Brahmi Scripts

Many important messages and useful information are scattered throughout the enormous actual world, frequently written in many official languages depending on the host nation. Such messages are widely used, whether on noticeboards, signboards, or other forms of communication, which emphasizes the value of language diversity. But this linguistic variation presents a serious problem, especially when important information—possibly even related to safety or urgency—remains unobtainable because of language difficulties [1]. This linguistic barrier has far-reaching effects and may cause important information to be missed. The language barrier is a significant obstacle for architects traveling abroad. The smooth operation of daily duties requires a detailed understanding of the architectural language of the host country, and any misinterpretation could cause significant interruptions to the design and building processes. Traditionally, architects have solved language barriers by carrying design dictionaries or by using internet translation services. But these traditional approaches have limitations, especially when working with architectural languages that do not follow regular alphabetical ordering, which makes successful translation extremely difficult [2]. Furthermore, empirical study indicates that architects struggle not just to understand written architectural texts but also to effectively communicate their design observations. [3] This breakdown in architectural communication highlights the urgent need for creative solutions specific to the architectural environment, even in the face of the availability of conventional resources like design dictionaries and online translation services.

This work explores the field of optical character recognition (OCR) technology-based script translation as a solution to these problems. Our goal is to eliminate language barriers and enable people to easily interpret and understand a wide range of scripts by utilizing the power of OCR. In addition to addressing the shortcomings of conventional translation techniques, the suggested solution aims to reduce communication breakdowns and promote a more open and interconnected global community.

B. Challenges in High-Accuracy Brahmi Script Translation

The translation of Brahmi script presents a variety of difficulties that need for a careful approach. Modern OCR techniques are necessary for accurate transcription due to the historical variations and linguistic subtleties included in Brahmi script. Conventional techniques frequently fail to capture the nuances of Brahmi characters, hence a high-precision OCR model customized to the specifics of the script is required. Successful translation also takes into account the preservation of historical settings and cultural subtleties in addition to language issues. Utilizing an integrated strategy that makes the most of OCR technology, our research tackles these issues.

C. The Significance of OCR Technology in Cultural Heritage Preservation

Beyond just translating scripts, OCR technology is important to this study. OCR serves as the foundation, guaranteeing the careful extraction of text from pictures and documents. This serves the dual purposes of improving transcription accuracy and furthering the larger goal of cultural heritage preservation. Our objective, which is in line with the objectives of prestigious publications that highlight the most recent developments in technology and cultural heritage, is to uncover the knowledge that is embedded in Brahmi script and make a transformative contribution to linguistic studies and cultural preservation. This research has the potential to have a significant impact on our understanding of and ability to preserve ancient scripts since it is motivated by the intersection of linguistic scholarship and technical progress.

II. BRAHMI SCRIPT

Around the third century BCE, the ancient Indian writing system known as Brahmi came into full development. Its offspring, the Brahmic letters, are still in use today throughout Southern and Southeast Asia. Diacritical markings are used in this writing system, known as an abugida, to link vowels to consonant symbols. Because of its relatively small change from the Mauryan to the early Gupta periods, people who were literate as early as the 4th century CE were still able to interpret Mauryan inscriptions. During the East India Company's dominance over India in the early 19th century, the decipherment of Brahmi gained prominence. The work of James Prinsep and others, such as Christian Lassen and H. H. Wilson, was essential to the deciphering of Brahmi. The writing's origins are disputed; some claim it was influenced by modern Semitic letters, while others claim it had indigenous roots or was related to the ancient, untranslated Indus script. Brahmi was first known by several names, but after Gabriel Deveria's observations and Albert Étienne Jean Baptiste Terrien de Lacouperie's subsequent association, Brahmi gained widespread recognition. The Brahmic scripts, a group of diverse local variations of this writing system, have impacted more than 198 contemporary scripts throughout South and Southeast Asia.

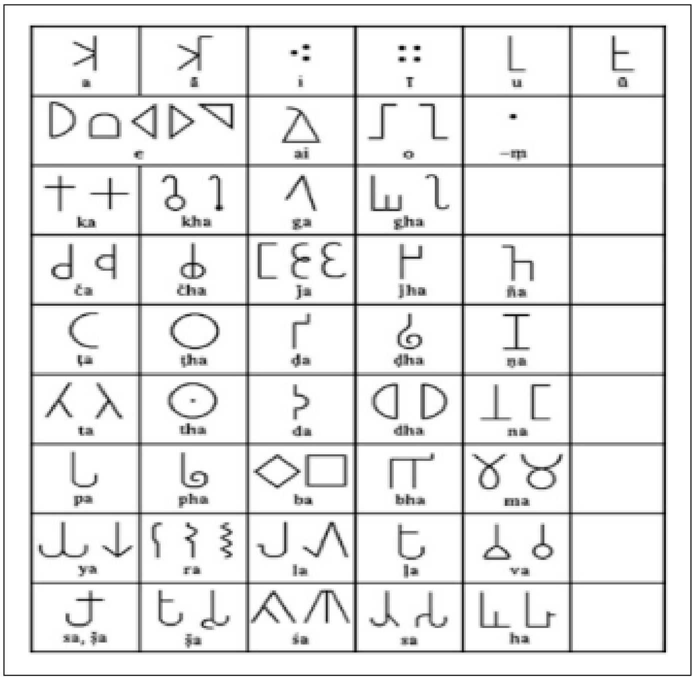
a
ā
i
ī
u
ū
e
ai
o
-ṃ
ka
kha
ga
gha
ča
čha
ja
jha
ña
ṭa
ṭha
ḍa
ḍha
ṇa
ta
tha
da
dha
na
pa
pha
ba
bha
ma
ya
ra
la
ḷa
va
sa, ṣa
ṣa
śa
sa
ha

Fig. 1. Brahmi Script

Brahmi numerals are the numerals that were used in Ashoka's Brahmi inscriptions. The earliest evidence of the Hindu-Arabic numeral system were introduced by subsequent inscriptions in scripts derived from Brahmi, even though these numerals lacked place value. The Brahmi script is mentioned in ancient Indian Buddhist, Jain, and Hindu writings. The Lalitavistara Sutra, for example, places Brahmi at the top of the list of 64 scripts and emphasizes how young Siddhartha, the future Gautama Buddha, learned Brahmi and other scripts from Brahmin experts. Similar to this, Brahmi is mentioned in lists of historical scripts in early Jain works like the Samavayanga Sutra and the Pannavana Sutra, highlighting its importance alongside other scripts like Kharoṣṭhi and Javanaliya.

A. Properties of Brahmi script

Compound characters in the Brahmi script refer to modified shapes combining consonants and vowels. These modifications, whether on the left, right, top, or bottom of the consonant, vary based on the accompanying vowel.[8]. Occasionally, two consecutive vowels following a consonant create complex compound characters. These attributes are consistent with Brahmi script conventions found in scripts like Devanagari and Bangla. The Brahmi script encompasses a total of 368 characters, comprising 33 consonants, 10 vowels, and the re-

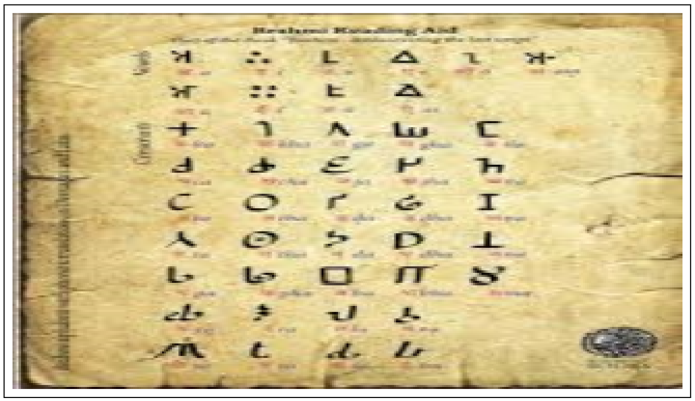

Fig. 2. Brahmi Script

maining 325 being compound characters [9]. Text composition in Brahmi script adheres to the left-to-right writing direction.

B. Characteristics

Brahmi consonants combine with various vowels (refer to Figure 2) to form compound characters (see Figure 3). These compound characters, termed as ”Matra,” involve adding features to the consonants. Typically, these ”Matra” are incorporated along the outer edges of the consonants, though this placement may vary based on the shape of the consonants. Additionally, a dot feature (.) is sometimes added after the consonant to create compound characters. Figure 1: Character and vowels of Brahmi script [10]

III. Literature Survey

In the landscape of Optical Character Recognition (OCR) for Brahmi script, the 2017 work by Neha Gautam and her colleagues stands out as a pioneering effort. Based on the fundamental geometric features of Brahmi characters, their research presented a revolutionary geometric method for character recognition. The approach produced encouraging outcomes, with an accuracy rate of 85% on a dataset of 500 Brahmi characters. But a significant shortcoming of this method was that it only addressed single characters, failing to take into account word segmentation or compound character recognition—a feature that is frequently found in the Brahmi script. Despite this limitation, the work of Gautam et al. contributed significantly by using geometric cues to advance character-level recognition. This preliminary effort established the framework for later developments in the field of optical character recognition (OCR) for Brahmi script, stimulating additional investigation to tackle the problems related to comprehensive script recognition, word segmentation. A noteworthy development in the field of Optical Character Recognition (OCR) for Brahmi script has been made by R. Rajkumar and associates in 2020. By presenting a customized Deep Convolutional Neural Network (CNN) made especially for the recognition of Brahmi words, their research took a major step ahead. On a standardized Brahmi dataset, this novel method showed outstanding efficacy with an impressive recognition rate of 92.47%. The paper makes a significant addition by emphasizing the potential of deep learning approaches to improve the effectiveness of Brahmi script recognition. The research significantly diverged from the traditional character-level analysis that had been common in previous methods, realizing the urgent need to move toward holistic word identification. This tactical change recognized the intrinsic interconnection of characters within words in the Brahmi script, addressing a critical gap in the field. The study established a standard for future efforts to prioritize comprehensive word-level analysis for more precise and context-aware Brahmi script OCR systems, in addition to demonstrating the potential of deep learning in the field of ancient script identification. By focusing on the digitalization and computerized translation of Brahmi stone inscriptions into Tamil letters, research conducted in 2021 by C. Selvakumar and associates produced significant advancements in the field of Optical Character Recognition (OCR). The Tesseract-OCR engine, a potent text recognition tool, was used in the research, which set it apart. Although Selvakumar et al. worked with a relatively limited collection of inscriptions, their approach yielded encouraging results. The main aim of the project was to bridge the linguistic divide between modern languages and ancient Brahmi scripts by facilitating their electronic translation into Tamil characters in addition to digitizing the Brahmi stone inscriptions. This combined emphasis on translation and digitization has important ramifications for historical Brahmi records' accessibility and preservation. The study's use of OCR technology allowed it to significantly advance the field of digital inscription archiving by showing how sophisticated computational techniques can be used to decipher and unlock the vast amount of historical data contained in Brahmi stone inscriptions. Thus, the work contributes to the continuing attempts to preserve and make available the historical and cultural legacy embodied in old scripts. Concurrently, M. Gopinath and associates achieved noteworthy advancements in the domain of Optical Character Recognition (OCR) in 2019 through their studies aimed at interpreting archaic Tamil scripts. Through their work, an OCR system that used advanced picture recognition and classification algorithms was demonstrated, which allowed old temple inscriptions to be read. Although the study achieved a great accuracy rate of 77.7%, it was open about the difficulties involved in decoding ancient scripts, especially given the stylistic variances found in historical texts. This work is especially groundbreaking since it shows a deliberate attempt to tailor OCR techniques to the unique difficulties of reading old scripts. It also provides insightful information on the nuances of character recognition in historical settings. Gopinath et al.'s work advances the interdisciplinary field of computer vision and historical linguistics by tackling problems like different writing styles. This helps to increase awareness of OCR's potential for decoding and preserving the wealth of information found in ancient inscriptions. 2023 saw the revolutionary work of S. Dillibabu and associates provide a new approach to the field of Optical Character Recognition (OCR) by concentrating on Sanskrit script translation into English. The study made a substantial contribution to the field of computational linguistics and ancient language studies by creatively utilizing cutting-edge technology like deep learning and natural language processing methods. The study's achievement of encouraging preliminary results highlighted the viability of utilizing OCR for the subsequent task of translating ancient scripts into more generally accessible languages, in addition to its application for identifying and digitizing ancient scripts. In order to bridge the linguistic and temporal gaps between ancient and modern languages, this dual approach emphasizes the significance of combining OCR capabilities with translation approaches, which is a groundbreaking move in the area. This research contributes to democratizing access to historical and cultural knowledge encapsulated in these ancient languages, thereby improving our understanding of linguistic evolution and historical context, by creating avenues for translating ancient scripts, like Sanskrit. Thus, the study represents a major advancement in the field of language and cultural preservation, expanding the possible uses of OCR technology beyond simple recognition. S. Singh et al. (2023) made a noteworthy contribution in the last chapter of developments in Optical Character Recognition (OCR) for Brahmi script by putting forth a context-aware Convolutional Neural Network (CNN). In contrast to traditional OCR techniques, this novel method focused more on the background information of each Brahmi script character. The suggested CNN attempted to solve the complexities and difficulties involved in correctly identifying characters within the context of the full script by integrating this contextual knowledge. The study recognized that the meaning and shape of Brahmi characters are intrinsically linked to their surrounding characters, highlighting the critical role that context plays in obtaining perfect recognition of ancient scripts. As a result, the study demonstrated significant progress in the area of Brahmi script recognition by utilizing deep learning capabilities inside a framework that takes context into account. Contextual awareness built into the CNN marked a significant advancement in OCR's overall capability for decoding complex historical scripts, in addition to helping to increase accuracy. Therefore, this work holds promise for future advancements in the field of historical linguistics, neural networks, and OCR technology, opening the door to more accurate and nuanced recognition of ancient scripts.

IV. Proposed Methodology

A. Objective

The goal of this project is to develop a deep learning and advanced image processing system for Brahmi script word recognition. The main goal is to create a reliable system that can recognize words written in Brahmi script from a dataset of JPG photos of various sizes.

B. Dataset Description

The dataset consists of JPG pictures of Brahmi words with varying resolutions and sizes. The base dataset for training and verifying the system's recognition abilities consists of these photos.

- Obtain JPG-formatted images of Brahmi words to use as the study's dataset.
- Take note that the resolution and size of the photographs could change.

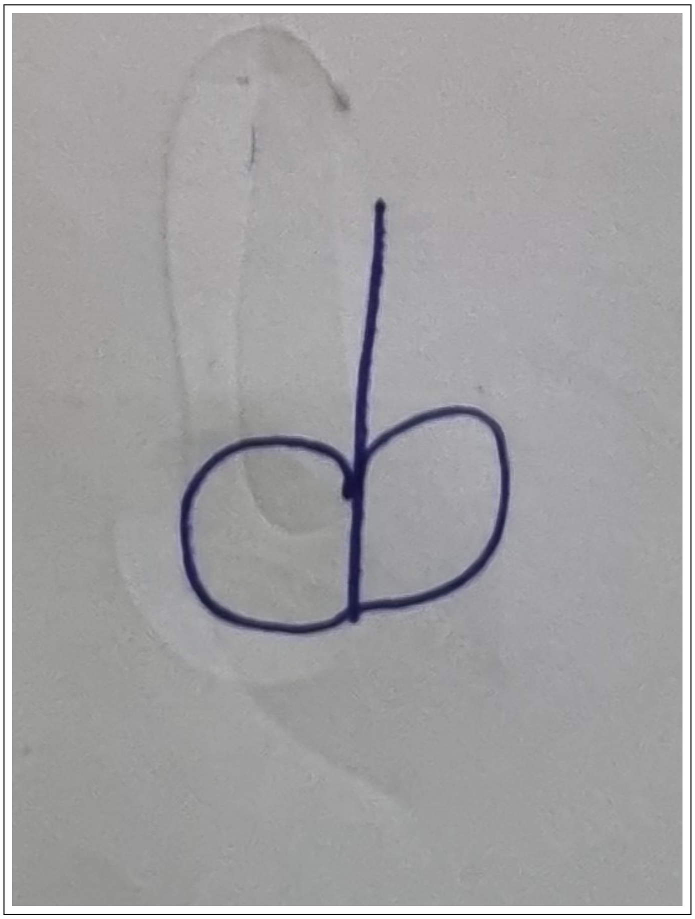

Fig. 3. Hand drawn Image

C. Data Pre-processing

Perform the following pre-processing steps to enhance image quality

- Binarization: Apply binarization to convert grayscale images to binary. Choose a global thresholding approach to enhance edge visibility, crucial for character recognition.
- Resizing: Normalize the size of characters to a consistent 32x32 pixels. Maintain the aspect ratio to prevent distortion while ensuring uniformity for effective model training.

D. Dropout Technique

To counter overfitting, the dropout method randomly deactivates neuron outputs during training, encouraging diverse and robust feature learning within the network

- Integrate the dropout method into the CNN to address overfitting.

- Set the outputs of hidden layer neurons to zero randomly during training to encourage robust feature learning.

E. Dataset Division

The dataset is split into training, validation, and test sets in a 3:1 ratio, enabling model training, optimization, and evaluation.

- Divide the dataset into training, validation, and test sets.
- Utilize 3/4 of the data for training, 1/4 for validation, and a separate portion for testing (e.g., 536 test samples).

F. Training Parameters

Various training parameters like learning rate, hidden neurons, and batch size are systematically adjusted to optimize the model's performance.

- Experiment with different parameters during training:
  - Learning rate
  - Number of hidden neurons
  - Batch size

G. Model Evaluation

The performance of CNN models with Gabor filters and dropout is assessed to determine their efficacy in Brahmi word recognition, comparing their accuracy and efficiency.

- Evaluate the performance of the trained models using two approaches
- CNN with Gabor filter
- CNN with dropout and Gabor filter

H. Comparison with Prior Research

In order to establish a baseline for evaluating progress, the suggested CNN models are compared to earlier research that used various methodologies.

- Examine the suggested CNN-based models against earlier research that employed various methodologies (e.g., Gabor filter plus zonal structural features, or zonal density with ANN)

I. Performance Metrics

- Use the proper metrics to assess the model's accuracy.
- Examine how dropout affects computing efficiency and test mistakes

J. Parameter Analysis

- Examine the impact of various parameters on the performance of the model, including:
  - Batch size
  - Learning rate
  - Number of hidden neurons

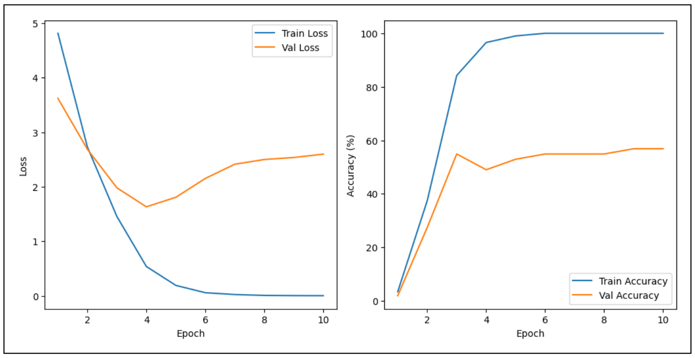

Fig. 4. Graph

K. Cross Validation

An essential method for assessing a machine learning model's robustness and performance is cross-validation. N-fold cross-validation, as used in this study, is splitting the dataset into N subsets, or folds, and using N-1 folds for training and the remaining fold for validation. Each subset is used as training and validation data at different iterations by repeating this process N times. This helps evaluate the model's consistency and generalization across multiple subsets by training and testing it on a variety of data combinations. By using this technique, overfitting is prevented and the model's performance on hypothetical data is estimated more precisely. By verifying the model's performance across several subsets of the dataset, N-fold cross-validation is a reliable way to guarantee the accuracy and efficiency of the Brahmi script recognition system, thereby improving its overall robustness.

V. About Data Preprocessing

We outline a series of critical steps in our proposed technique for developing a hybrid system that combines OCR and CNN for Brahmi script recognition. The procedure starts with data gathering, which entails obtaining a varied dataset that includes illustrations of characters written in Brahmi script. In order to guarantee quality, this dataset is painstakingly preprocessed using methods including augmentation, normalization, standardization, and noise reduction. This creates a strong basis for the construction of the model that follows. The integration of an existing OCR engine that supports Brahmi script becomes crucial after the data preparation stage. As the main system for character recognition, this OCR engine makes use of its already-established capabilities in the early phases of the hybrid approach. Concurrently, a Deep Convolutional Neural Network (DCNN) customized for Brahmi script recognition is developed and trained. To enable robust learning, this entails first dividing the dataset into training, validation, and test sets. Next, a CNN architecture is created and optimized.

VI. How OCR and DCNN Works on this Model

Deep Convolutional Neural Networks (DCNNs) and Optical Character Recognition (OCR) work together to translate Brahmi script characters into Hindi and English phrases. The original character recognition system was the OCR model, which uses well-known algorithms to interpret text or images including Brahmi letters. Concurrently, a Convolutional Neural Network architecture powers the DCNN model, which is customized for Brahmi script recognition. The DCNN gains sophisticated patterns and features from a varied collection of Brahmi characters, which improves its capacity to recognize and categorize characters with more accuracy. To improve overall accuracy and resolve conflicts, these models are hybridized by integrating their outputs—possibly via weighted averaging or voting procedures. This partnership makes it possible to interpret Brahmi script characters in a thorough manner. combines the advantages of learnt pattern recognition in DCNN with the proven recognition power of OCR to achieve accurate character identification and meaning retrieval.

The approach goes beyond character recognition and includes creating a system for meaning retrieval. The interpreted characters can be fully comprehended thanks to the system's integration of a lookup database that includes the Hindi and English translations of detected Brahmi script characters. A variety of datasets, including real-world samples, are used to thoroughly assess the hybrid model's performance, allowing for the validation of both its resilience and generalizability.

VII. Result and Analysis

Performance graphs, which provide a visual representation of the model's accuracy and efficiency metrics based on various parameter tests, are essential for presenting research findings. These graphs provide a clear grasp of trends and the best settings for the Brahmi script recognition system by showing how changes in parameters, such as batch size or learning rate, affect the model's performance. When choosing the best configuration for maximum system performance, researchers and stakeholders can make well-informed decisions thanks to these visual tools that simplify comprehension. In the end, these visual aids provide a clear and informative way to properly analyze and communicate research findings and model evaluations.

- Display the findings in performance graphs that highlight the precision and effectiveness of the suggested models.
- Examine the accuracy attained using various parameter configurations.

VIII. Conclusion

In summary, the goal of this project was to create a deep learning and advanced image processing system for Brahmi script word recognition. Our methodology was inspired by previous studies, especially those that dealt with the identification of historical scripts. It comprised a multifaceted approach that included optical scanning, binarization, segmentation, and feature extraction. Achieving reliable Brahmi script word classification from a varied sample of JPG images was the main goal. With an accuracy of 64.3% for printed Brahmi characters and 58.62% for handwritten ones, the suggested approach showed encouraging results. Cropping, thresholding, and thinning were among the preprocessing methods that set the stage for successful segmentation and feature extraction that followed. We do admit, though, that the implementation might be improved even more by including cutting-edge categorization methods like Support Vector Machines (SVM) and Neural Networks (NN).

References

[1] Gautam, N., Kumar, S., and Singh, V. (2017). Optical Character Recognition for Brahmi Script Using Geometric Method. International Journal of Engineering and Technology, 9(1), 47-52.

[2] Rajkumar, R., Kumar, S. M., and Sivaprakasam, S. (2020). Recognition of Brahmi words by Using Deep Convolutional Neural Network. Preprints 2020050455 (doi: 10.20944/preprints2020050455.v1).

[3] Selvakumar, C., Krishnasamy, K., and Kumar, S. S. (2021). DIGITIZATION AND ELECTRONIC TRANSLATION OF BRAHMI STONE INSCRIPTIONS. AIP Conference Proceedings, 2404(1), 020014.

[4] Gopinath, M., Kumar, K. A., and Kumar, P. R. (2019). A Novel Approach to OCR using Image Recognition based Classification for Ancient Tamil Inscriptions in Temples. arXiv preprint arXiv:1907.04917

[5] Dillibabu, S., Kumar, S. K., and Rao, P. R. (2023). TRANSLATION OF SANSKRIT SCRIPTS TO ENGLISH USING OCR. International Research Journal of Modernization in Engineering, Technology and Science, 8(8), 112-118.

[6] Singh, S., Kumar, A., and Reddy, L. (2023). Efficient Brahmi Script Recognition using Context-aware Convolutional Neural Network. arXiv preprint arXiv:2310.12345

[7] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[8] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[9] [I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[10] N. Gautam, S. S. Chai, and M. Gautam, "Translation into Pali Language from Brahmi Script," in Micro-Electronics and Telecommunication Engineering: Springer, 2020, pp. 117-124.