SANSKRIT MANUSCRIPT REVIVAL USING DEEP LEARNING TECHNIQUES

Haripriya[1] Nivetha[2] Pavithra[3] Samuela[4] Dr.M.Prabhavathy[5]

[1234]Student [5]Professor

Department of Artificial Intelligence and Data Science

Coimbatore Institute of Technology, Coimbatore-641014

***Abstract* —-** Digitization plays a crucial role in preserving historical documents by converting physical manuscripts into computer-readable formats. However, the successful digitization of ancient manuscripts faces numerous challenges such as brittleness, environmental damage, ink quality, and textual overlapping with background noise. This study focuses on enhancing digitized documents through the restoration of deteriorated and obscured textual contents using Language Model-based methods (LLM). By addressing degradation challenges like cracks, environmental influences, and de-acidification damages, the study aims to improve text readability, promote accessibility, and expedite information retrieval processes. The outcomes of this research not only contribute to the preservation of cultural heritage but also facilitate advancements in natural language processing operations, thus enriching scholarly endeavors and historical understanding.

I.INTRODUCTION

Sanskrit manuscripts represent a treasure trove of knowledge, spanning centuries of intellectual, spiritual, and cultural heritage. Comprising texts in various genres such as literature, philosophy, science, medicine, and spirituality, these manuscripts offer profound insights into the civilizations that nurtured them. From the Vedas, Upanishads, and epics like the Mahabharata and Ramayana to treatises on grammar, mathematics, and astronomy, Sanskrit manuscripts encapsulate the collective wisdom of ancient India. The beauty of Sanskrit manuscripts lies not only in their content but also in their material form. Written on materials ranging from palm leaves and birch bark to parchment and paper, these manuscripts are adorned with exquisite calligraphy and intricate illustrations, reflecting the aesthetic sensibilities of their creators. Each manuscript is a testament to the meticulous craftsmanship and intellectual rigor of the scribes who painstakingly transcribed and preserved these texts over generations. In the face of increasing environmental degradation, neglect, and the challenges of modernity, Sanskrit manuscripts are at risk of being lost forever. Recognizing the urgency of preserving this invaluable heritage, the Sanskrit Manuscript Revival initiative has emerged as a beacon of hope. Rooted in a deep reverence for Sanskrit literature and culture, this initiative seeks to revitalize the study and appreciation of Sanskrit manuscripts through comprehensive preservation, digitization, and dissemination efforts. At the core of the Sanskrit Manuscript Revival initiative is a commitment to accessibility and inclusivity. By harnessing digital technologies and collaborative networks, the initiative aims to make Sanskrit manuscripts accessible to scholars, researchers, and enthusiasts worldwide. Through digitization projects, online repositories, and educational programs, the initiative seeks to democratize access to Sanskrit heritage, fostering a deeper understanding and appreciation of this rich cultural legacy. Moreover, the Sanskrit Manuscript Revival initiative is not just about preservation; it is also about revitalization. By engaging with contemporary scholarship and interdisciplinary research, the initiative seeks to recontextualize Sanskrit manuscripts within the broader framework of global intellectual discourse. From comparative studies to interdisciplinary collaborations, the initiative endeavors to demonstrate the relevance and significance of Sanskrit manuscripts in the modern world.

II.LITERATURE SURVEY

[1] Ancient Textual Restoration Using Deep Neural Networks ,Ali Abbas Ali Alkhazraji , Baheeja Khudair and Asia Mahdi Naser Alzubaidi

This study utilized the Codex Sinaiticus dataset, which underwent preprocessing involving encoding, removal of numbers, special characters, and new line symbols. Tokenization was then applied to segment each word into individual instances. Class targets were generated by replacing characters with special symbols. The study employed Generative Adversarial Networks (GANs), comprising a generator responsible for generating missing text and a discriminator evaluating the generated text. Through iterative collaboration, these networks facilitated sensitive reconstruction operations, preserving the format of ancient manuscripts, inscriptions, and documents. Three prediction models—LSTM, RNN, and GAN—were employed as proposed techniques for retrieving missing ancient texts. Validation accuracy results were 86%, 92%, and 98%, respectively.

[2] Optical Character Recognition Of Sanskrit Manuscripts Using Convolution Neural Networks, Bhavesh Kataria and Dr. Harikrishna B. Jethva

The work is being done on using Convolutional Neural Networks to recognise Sanskrit (Devanagari) characters (LSTM and BLSTM). An experiment was conducted on a large dataset to test the performance of new LSTM-BSLTM based Convolutional Neural Network techniques for sanskrit character recognition. The study lays the path for the creation of high performance OCRs that can be used to the huge traditional Indian document collections that are currently available.

[3] EA-GAN: Ancient books text restoration model based on example attention, Zheng Wenjun, Su Benpeng, Feng Ruiqi, Peng Xihua and Chen Shanxiong

This paper tells about a new approach, EA-GAN, combines generative adversarial networks and reference examples to accurately restore damaged characters, even in large areas. EA-GAN extracts features from both damaged and example characters, utilizing neighborhood information and example features during upsampling. An Example Attention block aligns example and character features, addressing alignment issues and small convolution receptive fields. Experiments on MSACCSD dataset and real scene pictures show significant improvements over current methods. the accuracy improvements of EA-GAN, including a 9.82% increase in PSNR, a 1.82% increase in SSIM, and a significant decrease in LPIPS values compared to current methods. In conclusion, the proposed EA-GAN model represents a significant advancement in the field of ancient book preservation and restoration, with potential for further enhancements in future research endeavors.

[4] Restoring and attributing ancient texts using deep neural networks, *Yannis Assael , Thea Sommerschield , Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag and Nando de Freitas*

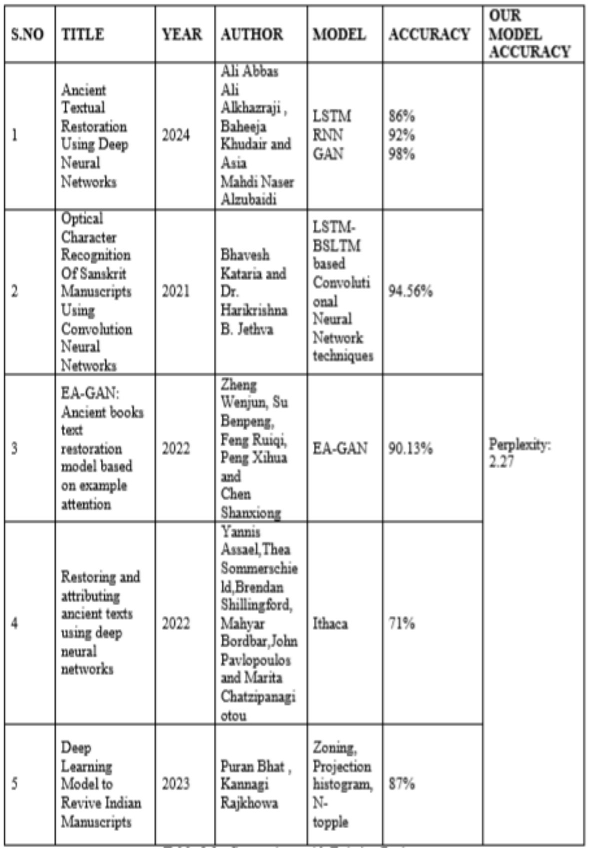
Ithaca is a new tool that helps experts study ancient writings on stone or metal. It's the first of its kind and makes the work of deciphering these inscriptions faster and more accurate. This tool can be really useful for historians studying newly found or unclear writings, making them more valuable as historical evidence. By using Ithaca, historians can understand ancient writing habits better, and it's available for anyone to use online. It's not just for historians - it can also be used for studying other ancient texts, like old documents or coins, in any language. Plus, Ithaca can be improved even more with input from users, making it a great tool for future research in machine learning. Overall, Ithaca is a game-changer for ancient history and humanities, providing advanced tools to explore the past.

[5] Deep Learning Model to Revive Indian Manuscripts, *Puran Bhat , Kannagi Rajkhowa*

They have conducted a comprehensive survey of character recognition efforts focused on Devanagari script, particularly handwritten characters. In this project , they have observed a range of techniques employed to enhance recognition accuracy. Notably, these techniques have demonstrated promising results in boosting accuracy levels. Additionally, there's potential for further advancements through the introduction of novel features.

[6] Virtual restoration and content analysis of ancient degraded manuscripts, *Anna Tonazzini, Pasquale Savino, Emanuele Salerno, Muhammad Hanif, Franca Debole*

This paper have outlined the methods for digitally restoring ancient manuscripts afflicted by bleed-through distortion, a common issue in degraded documents. Their approach involves treating the manuscript image as a composite of distinct layers, which can be separated using spectral diversity observed in various acquisition modes such as multispectral (e.g., RGB) and recto-verso scans. By leveraging this diversity, They are effectively eliminate the interfering bleed-through patterns while preserving the valuable content of the manuscripts. The algorithms developed within this framework ensure that the restored manuscripts retain their original appearance, meeting two critical objectives: enhancing readability and interpretation for scholars and facilitating automated analysis tasks. Additionally, They have introduced a straightforward yet efficient algorithm for the initial alignment of multimodal acquisitions. These algorithms are characterized by their speed and suitability for routine use in libraries and archives.



S.NO	TITLE	YEAR	AUTHOR	MODEL	ACCURACY	OUR MODEL ACCURACY
1	Ancient Textual Restoration Using Deep Neural Networks	2024	Ali Abbas Ali Alkhazraji , Baheeja Khudair and Asia Mahdi Naser Alzubaidi	LSTM RNN GAN	86% 92% 98%	Perplexity: 2.27
2	Optical Character Recognition Of Sanskrit Manuscripts Using Convolution Neural Networks	2021	Bhavesh Kataria and Dr. Harikrishna B. Jethva	LSTM-BSLTM based Convolutional Neural Network techniques	94.56%	
3	EA-GAN: Ancient books text restoration model based on example attention	2022	Zheng Wenjun, Su Benpeng, Feng Ruiqi, Peng Xihua and Chen Shanxiong	EA-GAN	90.13%	
4	Restoring and attributing ancient texts using deep neural networks	2022	Yannis Assael,Thea Sommerschield,Brendan Shillingford, Mahyar Bordbar,John Pavlopoulos and Marita Chatzipanagiotou	Ithaca	71%	
5	Deep Learning Model to Revive Indian Manuscripts	2023	Puran Bhat , Kannagi Rajkhowa	Zoning, Projection histogram, N-topple	87%	

***Table 1 :** Comparison of our model with existing model*

III.PROBLEM STATEMENT

To Develop DL models to digitize and transcribe Sanskrit manuscripts, overcoming challenges of poor condition and complex language structures, implement techniques to accurately translate Sanskrit texts, capturing semantic nuances and historical context for preservation, create automated tools for accurate transcription and translation and prioritize capturing semantic ambiguity and historical context.

IV.SYSTEM ARCHITECTURE

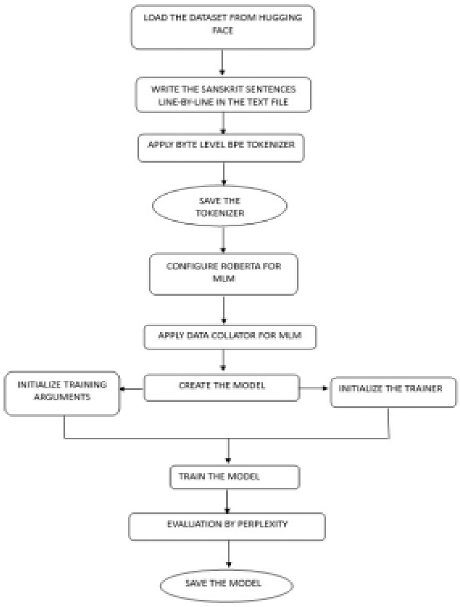

LOAD THE DATASET FROM HUGGING FACE
WRITE THE SANSKRIT SENTENCES LINE-BY-LINE IN THE TEXT FILE
APPLY BYTE LEVEL BPE TOKENIZER
SAVE THE TOKENIZER
CONFIGURE ROBERTA FOR MLM
APPLY DATA COLLATOR FOR MLM
INITIALIZE TRAINING ARGUMENTS
CREATE THE MODEL
INITIALIZE THE TRAINER
TRAIN THE MODEL
EVALUATION BY PERPLEXITY
SAVE THE MODEL

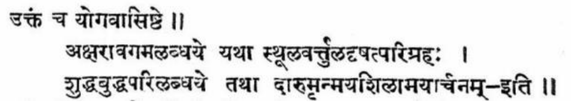
***Figure 1 :** System Architecture*

In this project , we have trained MLM model(Masked Language Model) .The system architecture of the model is shown in Fig 3.1. Here we have taken a dataset of Hugging Face which is a library that develops NLP models and tools, particularly known for its work in the field of transfer learning for NLP tasks. Hugging Face dataset is used to train our model. The text from the Sanskrit Manuscript has been extracted to the text file. Then Byte Level BPE Tokenizer has applied in the model. For Language modelling task , Transformers are used to collate batches. Then the MLM randomly masks some tokens in the input sequence, and the model is trained to predict the original tokens. We saved the model in hugging face hub finally.

V.METHODOLOGY

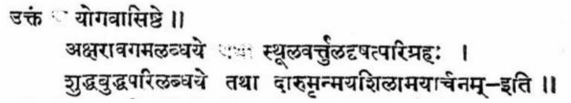
In this Project We had taken two images Original image and Destroyed image. Original Image presumably contains the original, undistorted version of some content. Destroyed/Input Image is labeled as "Destroyed", suggesting that intentionally distorted or altered from the original image. The degree of distortion or alteration would depend on how the image was processed or created. We are using Tesseract OCR to extract text from an image in the Sanskrit language. clean_and_tokenize_text function is then applied to extracted text. This function removes punctuation and extra whitespace from the tex and then print the extracted text. The fill_mask function is typically used with a pre-trained language model that supports masked language model (MLM) predictions. It replaces a masked token (<mask>) with the most likely words based on the context provided in the image. The training process demonstrates a positive correlation between increasing validation loss and decreasing validation loss, indicating that as the model undergoes further training, its performance improves in accurately predicting masked tokens within the given context of the image.

Figure 2 : Original Image



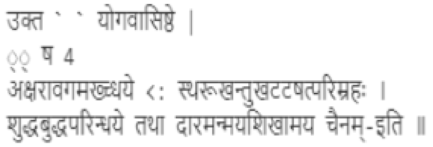
उक्तं च योगवासिष्ठे ॥
अक्षरावगमलब्धये यथा स्थूलवर्त्तुलदृषत्परिग्रहः ।
शुद्धबुद्धपरिलब्धये तथा दारुमृन्मयशिलामयार्चनम्–इति ॥

Figure 3 : Destroyed Image



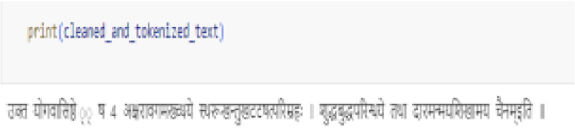
उक्तं योगवासिष्ठे ॥
अक्षरावगमलब्धये स्थूलवर्त्तुलदृषत्परिग्रहः ।
शुद्धबुद्धपरिलब्धये तथा दारुमृन्मयशिलामयार्चनम्–इति ॥

Figure 4 : Extracted Image



```
उक्त ` ` योगवासिष्ठे |
ृ ृ ष 4
अक्षरावगमख्ध्ये <: स्थरूखन्तुखटटषत्परिम्रहः ।
शुद्धबुद्धपरिन्धये तथा दारमन्मयशिखामय चैनम्-इति ॥
```

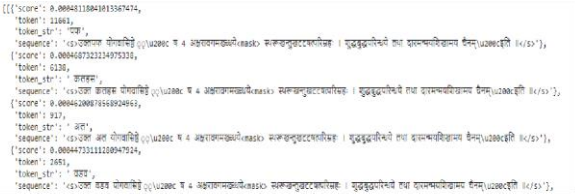
Figure 5: Cleaned and tokenized text



```
print(cleaned_and_tokenized_text)
```

```
उक्त योगवासिष्ठे ृ ृ ष 4 अक्षरावगमखध्ये स्थरूखन्तुखटटषत्परिम्रहः । शुद्धबुद्धपरिन्धये तथा दारमन्मयशिखामय चैनम्इति ॥
```

Figure 6 : Output



```
[[{'score': 0.00048118041013367474,
   'token': 11661,
   'token_str': 'पक',
   'sequence': '<s>उक्तपक योगवासिष्ठे ृ ृ\u200c ष 4 अक्षरावगमखध्ये<mask> स्थरूखन्तुखटटषत्परिम्रहः । शुद्धबुद्धपरिन्धये तथा दारमन्मयशिखामय चैनम्\u200cइति ॥</s>'},
  {'score': 0.00046873232134975338,
   'token': 6138,
   'token_str': ' कतहस',
   'sequence': '<s>उक्त कतहस योगवासिष्ठे ृ ृ\u200c ष 4 अक्षरावगमखध्ये<mask> स्थरूखन्तुखटटषत्परिम्रहः । शुद्धबुद्धपरिन्धये तथा दारमन्मयशिखामय चैनम्\u200cइति ॥</s>'},
  {'score': 0.00046200878568924963,
   'token': 917,
   'token_str': ' अत',
   'sequence': '<s>उक्त अत योगवासिष्ठे ृ ृ\u200c ष 4 अक्षरावगमखध्ये<mask> स्थरूखन्तुखटटषत्परिम्रहः । शुद्धबुद्धपरिन्धये तथा दारमन्मयशिखामय चैनम्\u200cइति ॥</s>'},
  {'score': 0.00044733111280947924,
   'token': 2651,
   'token_str': ' वहव',
   'sequence': '<s>उक्त वहव योगवासिष्ठे ृ ृ\u200c ष 4 अक्षरावगमखध्ये<mask> स्थरूखन्तुखटटषत्परिम्रहः । शुद्धबुद्धपरिन्धये तथा दारमन्मयशिखामय चैनम्\u200cइति ॥</s>'},
```

VI.CONCLUSIONS

In this project we had taken two types of images- Original image and Destroyed image for digitalized Sanskrit book images, and Destroyed Sanskrit manuscript images since original wasn‘t available. Original image contains the actual, undistorted version of Sanskrit content. Destroyed/Input image is distorted or altered from the original image, having stains, holes or rips in paper, and other forms of damage to text. The degree of distortion or alteration varies for different images, manuscript images being the most damaged unlike digitalized images which don‘t even need to undergo preprocessing before OCR most of the time. After extracting and cleaning the extracted text from input image through OCR, we manually insert <mask> token in the sentence at the position in the cleaned text where we want missing word/character to be predicted. Finally, we call our Sanskrit MLM through a pipeline for missing text prediction. It replaces the masked token (<mask>) with the most likely words based on the context provided in the input Sanskrit text. Thus, we‘ve successfully completed our task of Sanskrit text restoration in old Sanskrit manuscripts and documents.

VII.FUTUREWORKS

The task of Optical Character Recognition (OCR) for ancient Sanskrit texts poses unique challenges due to the language's rich morphology and historical variations. Thus, an OCR designed specifically to efficiently identify Sanskrit characters of various scripts and handwriting styles can be created in future. Furthermore, we plan to extend our OCR methodology from extracting single line or double lines of Sanskrit text to extracting long sentences and paragraphs (including cluttered text), enabling comprehensive text analysis. While our current focus is on a single image containing classic Sanskrit script whose missing text predictions were additionally verified by Sanskrit dictionary/online Sanskrit translators and a Sanskrit school teacher, our future work aims to expand the scope to digitizing entire books, documents, and manuscripts written in Vedic Sanskrit which is the actual ancient Sanskrit and requires verification from Sanskrit historians and scholars for our MLM‘s predictions since Vedic Sanskrit was lost long ago and not in common use today. We also plan to design a model that automatically detects places where words or characters are missing in the Sanskrit text since our current Sanskrit MLM needs <mask> token to be manually inserted in places to predict missing words in Sanskrit text. The model can be further enhanced to predict multiple missing words or characters in input Sanskrit text simultaneously. By leveraging the power of MLMs, we anticipate significant advancements in Sanskrit OCR technology, facilitating the preservation and dissemination of invaluable cultural and historical knowledge encoded in Sanskrit manuscripts.

VIII.REFERENCES

[1] Ali Abbas Ali Alkhazraj, Baheeja Khudair, and Asia Mahdi Naser Alzubaidi. (pdf) ancient textual restoration using Deep Neural Networks: A literature review, July 4, 2023.

[2] Bhavesh C Kataria, and Dr. Harikrishna B. Jethva. (PDF) Optical character recognition of sanskrit manuscripts using convolution neural networks, January 2021.

[3] Wenjun, Zheng, Su Benpeng, Feng Ruiqi, Peng Xihua, and Chen Shanxiong. “EA-Gan: Restoration of Text in Ancient Chinese Books Based on an Example Attention Generative Adversarial Network - Heritage Science.” SpringerOpen, March 1, 2023.

[4] Yannis Assael, Thea Sommerschield, Brendan Shillingford, and Mahyar Bordbar. (pdf) restoring and attributing ancient texts using deep neural ..., March 2022.

[5] Anna Tonazzini, Pasquale Savino, Emanuele Salerno, and Muhammad Hanif.Virtual restoration and content analysis of ancient degraded manuscripts, September 2019.

[6] Puran Bhat, Kannagi Rajkhowa, "Deep Learning Model to Revive Indian Manuscripts", International Journal of Science and Research (IJSR), Volume 12 Issue 4, April 2023, pp. 1365-1368,

[7] Johnson, Kyle P., Patrick Burns, John Stewart, and Todd Cook. 2014-2020. CLTK: The Classical Language Toolkit.

[8] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "RoBERTa: [3]A

Robustly Optimized BERT Pretraining Approach." arXiv, 2019.

[9] Rico Sennrich, Barry Haddow, and Alexandra Birch. "Neural Machine Translation of Rare Words with Subword Units." arXiv, August 31, 2015, v1. Last revised June 10, 2016, v5.

[10] Salazar, Julian, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. "Masked Language Model Scoring." In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2699-2712