Beyond Digital Boundaries: Examining Record Keeping Practices Through Handwritten Annotations

Gargy G, *Department of Computer Science and Engineering, Noorul Islam Centre for Higher Education,*
A. Shajin Nargunam, *Director-Academics, Department of CSE, Noorul Islam Centre for Higher Education*

***Abstract*— Handwritten Malayalam Script Recognition (HMSR), an Indic language predominantly spoken in the Indian state of Kerala, presents a unique set of challenges due to the complex nature of its script. This script, derived from ancient Brahmi, features a combination of curved and straight lines, intricate ligatures, and a wide range of character set that make it a formidable task for automated recognition systems. In this abstract, we are mentioning the difficulties associated with recognizing handwritten Malayalam script and discuss the key approaches used in tackling this problem. This Systematic Review (SR) involved the gathering, consolidation, and examination of research articles pertaining handwritten recognition of the Malayalam script in publications spanning from 2006 to 2023. We conducted a systematic review, adhering to a predefined protocol to gather information from well-established electronic databases. We employed various search methods, including keyword searches, forward and backward reference searching, to comprehensively identify articles relevant to the topic. After a rigorous selection process, we identified and included 121 articles in this systematic review. This review article aims to present the latest findings and advancements in the field of Malayalam handwritten recognition while also shedding light on areas of research that require further exploration.**

***Index Terms*— Classification, Feature Extraction, Handwritten Malayalam script recognition.**

I. INTRODUCTION

HANDWRITTEN Malayalam script is renowned for its frequent utilization of conjunct characters where two or more base characters are combined into a single ligature. This complexity adds a layer of difficulty in recognizing and segmenting characters. It is a non-linear script, meaning characters can be written horizontally from left to right and also from top to bottom within a single word. This adds complexity to layout analysis and character segmentation.

The script contains intricate ligatures, where the shapes of characters change when they combine [1]. Recognizing and separating these ligatures is a significant challenge. Many characters in Malayalam have different forms depending on their position in a word. For instance, the same character may look different when it's in the initial, medial, or final position in a word. It uses diacritic marks to represent vowels. These signs can appear above, below, before, or after a consonant, making recognition more challenging [2].

Some characters in Malayalam are relatively complex, involving loops, curves, and intricate strokes [3]. Accurately identifying these characters is a complex task. It follows various contextual rules where character forms and ligatures change depending on the surrounding characters. Recognizing characters within their contextual context is intricate [4]. The script may contain characters with similar appearances; leading to recognition ambiguity. For example, the characters for "ra" and "la" can be visually similar. Different writers may use variations in script style and character shapes, adding to the complexity of recognition [5].

Malayalam is a syllabic script, which means each character represents a syllable rather than a single phoneme [8]. This further complicates recognition. To address these complexities, handwritten recognition systems for Malayalam script typically employ advanced techniques such as deep learning models for example Recurrent Neural Network (RNN) [9] and Convolutional Neural Network (CNN) [10] contextual analysis, and training on diverse datasets. They also make use of language models specific to Malayalam, taking into account the script's rules and variations. Despite its intricacies, the recognition of handwritten Malayalam script is essential for preserving cultural and historical documents and enabling digitization efforts in the language.

II. METHODS OF EVALUATION

As mentioned previously, this Systematic Review (SR) is conducted with the objective of identifying and presenting relevant literature pertaining to Handwritten Malayalam script recognition (HMSR). In conclusion, the review seeks to accomplish the following goals:

1) Summarize existing research efforts, encompassing databases and machine learning methods, in the field of handwritten Malayalam script recognition system.
2) Highlight areas of weakness within the existing research, with the intention of suggesting opportunities for further investigation and improvement.
3) Identify emerging research domains and directions within the broader of handwritten Malayalam script recognition.

Subsequent subsections will elaborate on the protocol for the review, methods for extracting and synthesizing data, methodology for searching and process of selection.

A. Protocol for the Review

Adhering to the principles and guidelines of a Systematic

Review (SR), this comprehensive investigation commenced by creating a detailed review protocol. The protocol delineates multiple elements, encompassing the review's context, search methodology, data extraction methods, research inquiries, and criteria for evaluating study quality and conducting data analysis [10].

It is worth emphasizing that the review protocol serves as a distinguishing factor between an SR and a traditional or narrative literature review. Moreover, it contributes to maintaining consistency throughout the review process and helps mitigate potential biases among researchers. This is primarily because researchers are required to provide a clear search methodology and establish criteria for data extraction methods and synthesizing data within the review.

B. Methods for Extracting and Synthesizing Data

Developing procedures for extracting and amalgamating data is a vital step in guaranteeing the inclusion of pertinent articles in the study. Our criteria encompass research studies published in journals, conferences, symposiums, books, dissertations and workshops focused on Handwritten Malayalam Script Recognition (HMSR). This Systematic Review (SR) specifically considered studies published between 2006 and 2023.

Initially, our keyword-based search yielded a total of 135 articles pertaining to handwritten Malayalam languages (refer to Fig.1 for an overview of how the selection process is done). After rigorously examining these articles, we excluded those that, despite matching keywords, were not closely associated with the handwritten Malayalam language. Furthermore, articles were omitted due to issues like duplication, unavailability of full text, or when they did not correspond with our research inquiries, these articles were excluded.

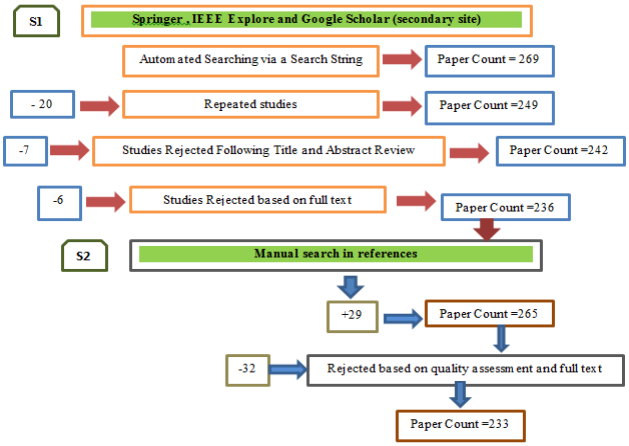

Fig.1. A comprehensive description of the criteria for selecting and incorporating studies into the review process.

C. Methodology for Searching

The method of searching strategy encompasses both automated and manual components, as illustrated in Fig.1. The automated search was instrumental in identifying primary studies and providing a broad perspective. To ensure a comprehensive review, we expanded our scope by incorporating additional studies. We applied a manual search technique to the references of studies that were discovered during the automated search.

For the automated search, we utilized established databases known for hosting pertinent research papers. These databases comprise Web of Science, Springer, IEEE Explore and Scopus (Elsevier). Although there is an abundance of literature available in newspapers, magazines, and blogs, However, these sources were not encompassed in this review due to their content not being subjected to a review process, making the quality verification unreliable.

In our search, we employed commonly used keywords derived from our research inquiries and study titles to discover research papers. The objective was to identify as many relevant papers to the greatest extent feasible from our primary collection of keywords. We explored various permutations of handwritten Malayalam concepts, including phrases like "Malayalam handwritten script recognition [1]," "pattern recognition in handwritten Malayalam scripts [3] ," and "offline Malayalam handwritten text recognition" ,"online Malayalam handwritten document" [5], "pattern matching in Malayalam handwritten documents", Malayalam handwritten segmentation techniques" [9] "Malayalam handwritten word recognition", "Malayalam handwritten palm leaf script recognition" [10] "Malayalam handwritten character recognition".

Following the initial phase of data collection through search queries, the subsequent phase involved analyzing the data to evaluate their alignment with the research inquiries and to establish methods for extracting and synthesizing data. To manage and store these related research papers for referencing, we utilized a Mendeley bibliography management tool. This tool also assisted in identifying and managing repeated studies. To ensure thoroughness, a manual searching was conducted in combination with the automated search, minimizing the possibility of overlooking relevant studies. This was accomplished by employing both backward and forward referencing.

For data extraction, all outcomes were brought into a Microsoft spread sheet. The snowballing technique, an iterative process verifying references in identified studies to find more pertinent research papers, was applied to initial phase of preliminary studies, further expanding the pool of relevant preliminary studies. The set of preliminary studies obtained after the snowball process was subsequently employed into Mendeley for reference and citation purposes.

D. Process of Selection

We utilized a staged approach for study selection. After conducting keyword searches across all relevant databases, we initially retrieved 269 research papers via automated search. Among these, 20 were identified as duplicates and were consequently removed. The remaining 249 studies underwent screening based on methods for extracting and synthesizing data, by considering their titles, abstracts, keywords, and publication types. This process resulted in the rejection of 7 studies, leaving us with 242 studies. Subsequently, the modes of selection were put into effect, leading to the rejection of an

additional 6 studies, resulting in the ultimate set of 236 studies.

Upon completing the automated search phase, we initiated a hand-search to ensure the comprehensiveness of our results of search. Then we meticulously screened the remaining 236 studies and scrutinized references to identify any relevant research papers that possibly could have been overlooked during the automated search. The manual search yielded an additional 29 studies. After incorporating these studies, a preparatory final list of 265 primary studies was compiled. Table 1 shows the papers undergone the selection process

TABLE 1. Presents the span of studies selected among various publications, after implementing the fore mentioned process of selection.

Journal	**Papers undergone selection process**
Springer	9
Elsevier	17
IEEE	42
Others	53
Total	121

E. *Quality Evaluation*

The last stage involved applying Quality Evaluation (QE) to the preparatory final list of 265 studies. The Quality evaluations were employed as the last step to refine the catalogue of studies identified for the Systematic Review (SR). QE typically identifies studies whose quality does not contribute effectively to answering the research questions. Following the application of QE, 32 studies were excluded, resulting in a final selection of 233 primary studies. From this 233 primary studies we have taken a total of 121 selected papers are chosen for final studies. Refer to Fig.1 for comprehensive, steps of overview of the selection process. Refer to Table 1 for the details of studies choosen after doing the selection process.

F. Extraction and Compilation of Data

In this phase, we collected metadata from the selected 121 studies. As mentioned earlier, Microsoft Excel, Mendeley and were employed for managing this metadata. The primary aim of this level was to document the information acquired from the preparatory initial set of studies. The data included the study number (for identification), study title, name of author, year of publication, publication platform (e.g., journals, conference proceedings, symposium, dissertation and books), citation counts, and details regarding the study's context (i.e., the techniques employed in the study). Data's where gathered following a comprehensive analysis of the study to find the techniques and algorithms implemented by the scholars. Table 2 displays the specific fields of extracted data from the research studies.

TABLE. 2 Obtained metadata attributes from the chosen research papers.

Chosen characteristics	**Illustration**
Study number	Distinctive identifier for the chosen research article.
Reference	Author, title, publication year etc
Category of document	Journal,workshop,conference, symposium,dissertation,book etc
Language	Malayalam
Citation	Count of references
Techniques Employed	Classification and Feature extraction techniques

III STATISTICAL FINDINGS OBTAINED FROM A CURATED SET OF STUDIES

Within this segment, we present the statistical outcomes of the chosen studies, categorizing them by their sources of publication, citation counts, Time-based perspective, languages used, and research techniques employed.

A. Overview of Publication Sources

In this comprehensive review, most of the research papers we've included have been published in reputable journals and prestigious conferences [20]. This suggests that, given the high quality of these research studies, our Systematic Review (SR) can serve as a valuable reference for identifying the latest trends and shedding light on potential research directions within the field of handwritten Malayalam scripts. Figure 3 illustrates the span of these studies across various publication sources.

Among the 121 studies incorporated into the analysis phase, a substantial 59 of them were featured in scholarly journals (representing 61%), while 55 studies were presented in conference papers (comprising 34%). On the other hand, a smaller number of 4 papers were found to be published in workshop, and only 1 relevant research papers were presented in symposium, 1 paper found in books and 1 in dissertation.

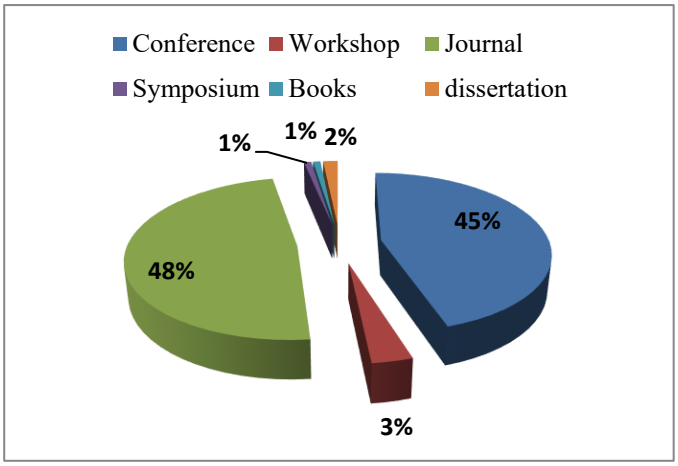

Fig. 2 Span of Studies by Publication Source

B. Citation of research

The citation counts were retrieved from Google Scholar, and, on the whole, the selected studies exhibited a good commendable citation record even though the researches under Malayalam handwritten scripts are in progress, which started in early 2006. This mirrors the high quality of the chosen studies, making them valuable additions to this review.

This also implies that ongoing research endeavors are in progress in this particular domain from 2017. As depicted in Figure 4, nearly 91% from the chosen studies with the exception

of one citation at minimum, except for the research articles published in 2023, which are relatively recent. Within our selection, 3 studies garnered over 100 citations, while 7 studies received citations ranging from 51 to 100 times. An additional 14 studies received citations in the range of 31 to 50 times, 18 studies comes under the range of 16 to 30 times, and 50 studies were cited between 1 and 15 times. (Refer to Fig.3 for an overview about the citation counts from the studies that were chosen).

Overall, we anticipate that the citation counts for the selected studies will continue to rise, given the on-going publication of research articles in this field. In Table 3, specific details about research publications and the citations can be found. These papers can be considered as having a significant impact on researchers striving to develop robust handwritten Malayalam Script recognition system.

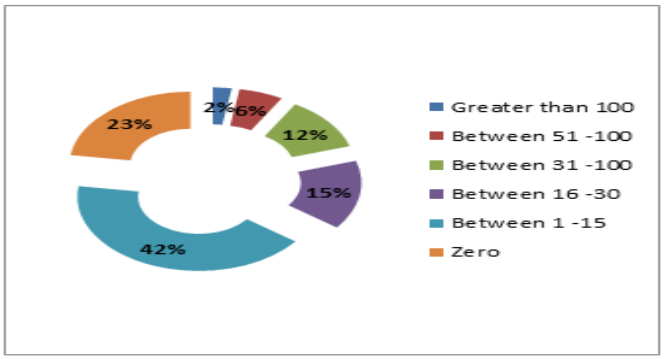

Fig. 3 The count of citations regarding the chosen studies

C. Time-based perspective

The span of studies from 2006 to 2023 is depicted in Fig.4. Referring to the figure, it is apparent that there is significant in the number of publications over the years. Notably, there existed a sharp increase in publications related to handwritten Malayalam script recognition in 2011 , 2017 , 2018 , and 2023 . From 2005 to 2010, the publication count remained relatively constant. Nonetheless, following 2010, there was a consistent upward trend, resulting in 99 publications over the 15-year period from 2006 to 2021. (Refer to Fig.4 for an outlook about the count of published works from 2006 - 2023)

Over the past two years, there was a significant surge, with 20 new studies, compared to 99 studies in the previous 15 years. This is unsurprising, given the growing interest of researchers in handwritten Malayalam script recognition due to advancements in computer vision and deep learning. It is anticipated that the application areas of handwritten Malayalam scripts will persist in the future.

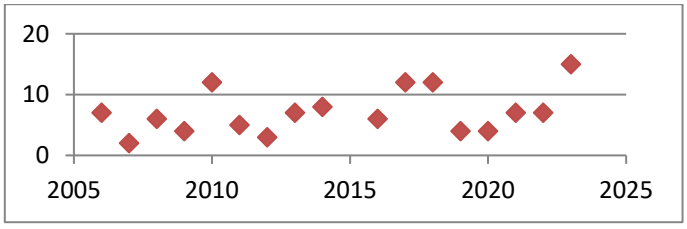

Fig. 4 Published works throughout the years. The y-axis represents the publication count

D. Current directions in research

Handwritten Malayalam Script Recognition (HMSR) research has seen a transition to deep learning methods in recent years. Although this approach has resulted in improved classification accuracy, it has come with enhanced computational demands, particularly during the the phase of training. In this section, we analyse trends in handwritten scripts in Malayalam research from 2006 to 2023 and present our findings in Table 5. This table provides an overview about the development of optical character recognition for handwritten Malayalam document, various techniques employed, publication years, and corresponding references.

Notably, a significant portion of recent publications has embraced deep learning techniques [8], particularly Convolutional Neural Networks [10], which is extensively utilized for hand-written Malayalam script recognition. This preference in the context of deep learning is often influenced by the presence of extensive datasets. making it effective in learning meaningful models. However, it's significant to observe that even as deep learning methods have enhanced classification accuracy. Though, this often comes with the trade-off of heightened computational complexity. Nonetheless, certain recent investigations have accomplished state-of-the-art outcomes by effectively integrating with the use of traditional feature extraction methods, alongside with feature selection algorithms.

V. DATA COLLECTION

A. GARGYG13

As a result of the less availability of Malayalam Handwritten dataset, we have created a fresh dataset for Malayalam named GARGYG13. It was originally created in 2023 by the scholars at Noorul Islam Study Centre for Higher Education India. The primary aim behind its progression was to advance the research and development of Malayalam handwriting script recognition systems. This dataset comprises a total of 1385 handwritten images. Within these images, all of which were written by 300 distinct writers, as depicted in Fig. 5. This dataset has been utilized for the effective recognition of Handwritten documents. https://dx.doi.org/10.5255/UKDA-SN-857993

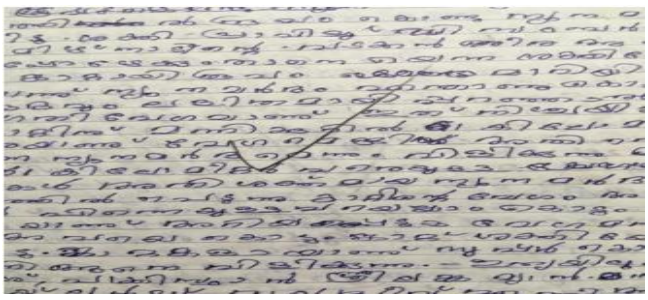

Fig. 5. Sample image from GARGY13 dataset

TABLE 3. Scholarly publications

Sl.No	Script	Techniques Employed in handwritten Malayalam Scripts	Year	Ref
1	Malayalam	HOG feature-based recognition	2018	[1]
2	Malayalam	Convolutional Neural Network (CNN) framework using alexnet	2018	[2]
3	Malayalam	Geometrical and Structural Properties	2018	[3]
4	Malayalam	Using deep learning approaches	2018	[4]
5	Malayalam	On developing handwritten character image database	2019	[5]
6	Malayalam	Residual Network enhanced by multi-scaled features	2019	[6]
7	Malayalam	Image Gradient approximations	2019	[7]
8	Malayalam	Deep architecture	2019	[8]
9	Malayalam	Multilevel CNN architecture	2020	[9]
10	Malayalam	Gabor based MultiLayer Architecture	2020	[10]
11	Malayalam	Text Line Extraction	2020	[11]
12	Malayalam	Deep Classification	2020	[12]
13	Malayalam	Spatial domain feature extraction	2021	[13]
14	Malayalam	Support Vector Machine	2021	[14]
15	Malayalam	Binarization using difference of concatenated convolutions	2021	[15]
16	Malayalam	A Dataset for Indic Handwritten Text	2021	[16]
17	Malayalam	Levenshtein distance based implementation	2021	[17]
18	Malayalam	Deep learning approach	2021	[18]
19	Malayalam	Preprocessing and Post processing Techniques from ancient palm leafs	2021	[19]
20	Malayalam	Global Semantic Information	2022	[20]
21	Malayalam	A review :Palm Leaf Malayalam Characters	2022	[21]
22	Malayalam	Transfer Learning Based Deep Neural Network	2022	[22]
23	Malayalam	Complete Denoising Solution	2022	[23]
24	Malayalam	latent dirichlet allocation and semantic features	2022	[24]
25	Malayalam	VGGNET and DENSENET using transfer learning approach	2022	[25]
26	Malayalam	Connected Component Analysis	2022	[26]
27	Malayalam	CNN Architecture	2023	[27]
28	Malayalam	Text Line Segmentation	2023	[28]
29	Malayalam	Character segmentation and Efficient feature extraction	2023	[29]
30	Malayalam	Transfer Learning and Fine Tuning of Deep Convolutional Neural network	2023	[30]
31	Malayalam	A review : Palm leaf manuscript	2023	[31]
32	Malayalam	Structural features using deep learning models.	2023	[32]
33	Malayalam	Malayalam palm leaf manuscript dataset	2023	[33]
34	Malayalam	Using Transfer Learning	2023	[34]
35	Malayalam	HOG-DCNN model with multiview augmentation and inference fusion	2023	[35]
36	Malayalam	Deteriorated image classification model	2023	[36]
37	Malayalam	Recognition for characters, digits and words: A comprehensive survey	2023	[37]
38	Malayalam	Competition on Indic Handwriting Text Recognition	2023	[38]
39	Malayalam	Binarization and Classification Using False Color Spectralization and VGG-16 Model	2023	[39]
40	Malayalam	A review :An overview of techniques and challenges	2023	[40]
41	Malayalam	Character Recognition and Genetic Algorithms	2023	[41]

Table 4. Inquiries for research and the driving force

Inquiries for Research	The driving force
What are the methods for extracting features and classifying them in handwritten Malayalam script recognition?	To discern patterns in the utilization of feature extraction methods and machine learning techniques spanning nearly two decades.
Which language is used for the investigation?	To shed light on the Malayalam languages that have been examined, thereby pinpointing those languages requiring further research attention.
What are the different data collections available for doing research?	Adequate data availability always constitutes a fundamental prerequisite for constructing handwritten Malayalam script recognition system
What are the emerging areas of research?	To offer direction for upcoming research endeavours.

VI. CURRENT DIRECTIONS IN RESEARCH

Handwritten Malayalam Script Recognition (HMSR) research has seen a shift towards deep learning methods in recent years. This method has led to improved classification accuracy, it has come with increased computational demands, especially during the training phase. In this section, we analyze trends in handwritten character recognition research from 2006 to 2023 and present our findings in Table 5. This table provides an overview about the development of handwritten Malayalam document, various techniques employed, publication years, and corresponding references.

Notably, a significant portion of recent publications has adopted deep learning methods, particularly Convolutional Neural Networks (CNN), which is extensively used for handwritten Malayalam script recognition. This preference for deep learning is often influenced by the availability of large datasets, making it effective in learning meaningful models. However, it's important to note that while deep learning methods have enhanced classification accuracy, they come at the cost of increased computational complexity. Nevertheless, some recent studies have accomplished state-of-the-art outcomes by effectively integrating traditional feature extraction methods with feature selection algorithms.

VII SUMMARY AND FUTURE PROSPECTS

A. Summary

1) Handwritten Malayalam Script Recognition technology was indeed developed several decades ago. These early systems were primarily designed for recognizing printed text and were often expensive and limited in accuracy.

Over the last few decades, the rise and widespread adoption of deep learning and machine learning methodologies have brought about a significant transformation in handwritten character recognition technology. These approaches have facilitated the creation of more precise and adaptable optical character recognition systems, some of which can now effectively identify handwritten text.

2) In our comprehensive literature review, we methodically gathered and examined research articles from 2006 – 2023 for handwritten Malayalam script. Our exploration revealed that certain techniques exhibited superior performance with Malayalam scripts, such as the convolutional neural network and transfer learning approaches. This variation can likely be attributed to how a particular technique models distinct character styles and the high quality of the dataset employed.

B. Future work

1) At government medical colleges in Kerala nearly 500 patients will visit every day for consultation, the implementation of Malayalam handwritten script recognition system help healthcare providers efficiently digitize and manage handwritten patient records in Malayalam. This can enhance data accuracy, reduce errors, and improve overall patient care. It can be employed for automatically extracting and processing handwritten prescriptions, ensuring that medications are dispensed accurately and reducing the potential for medication errors. This is crucial for patient safety.

REFERENCES

[1] Jayakumari, Bipin & Kavana, Amel. (2023). Classification of heterogeneous Malayalam documents based on structural features using deep learning models. International Journal of Electrical and Computer Engineering (IJECE). 13. 894. 10.11591/ijece.v13i1.pp894-901.

[2] B J, Bipin & Rani, N Shobha. (2023). HMPLMD: Handwritten Malayalam palm leaf manuscript dataset. Data in Brief. 47. 108960. 10.1016/j.dib.2023.108960.

[3] Jose, Bineesh & K P, Pushpalatha. (2023). Malayalam Handwritten Character Recognition Using Transfer Learning. 1-5. 10.1109/AICAPS57044.2023.10074586.

[4] Jose, Bineesh & K P, Pushpalatha. (2023). Classification of handwritten Malayalam characters using a HOG-DCNN model with multiview augmentation and inference fusion. Multimedia Tools and Applications. 1-12. 10.1007/s11042-023-16154-7.

[5] B J, Bipin & Rani, N Shobha & Khan, Mustaqeem. (2023). Deteriorated image classification model for malayalam palm leaf manuscripts. Journal of Intelligent & Fuzzy Systems. 45. 1-19. 10.3233/JIFS-223713.

[6] Singh, Sukhdeep & Sharma, Anuj & Chauhan, Vinod. (2023). Indic script family and its offline handwriting recognition for characters/digits and words: a comprehensive survey. Artificial Intelligence Review. 1-53. 10.1007/s10462-023-10597-y.

[7] Mondal, Ajoy & Jawahar, C.. (2023). ICDAR 2023 Competition on Indic Handwriting Text Recognition. 10.13140/RG.2.2.35969.33125. G.A., Jain, R., Kise, K., Zanibbi, R. (eds) .

[8] Nair, B & Raj, KV & Mangalath, Kedar & Pai, Vaishak & Sreejil, EV. (2023). Ancient Epic Manuscript Binarization and Classification Using False Color Spectralization and VGG-16 Model. Procedia Computer Science. 218. 631-643. 10.1016/j.procs.2023.01.045.

[9] Sarithadevi, R Rajesh - AIP Character recognition for Malayalam palm leaf manuscripts: An overview of techniques and challenges S Conference Proceedings, 2023 Volume 2773, Issue 1, id.020003, 5 pp.2023.DOI:10.1063/5.0138616

[10] Kasthuri, Magesh, and Vigneshwar Manoharan. "Handwritten Character Recognition and Genetic Algorithms." Handbook of Computational Sciences: A Multi and Interdisciplinary Approach (2023): 197-223.