OPEN ACCESS

Global Journal of Research in Engineering & Computer Sciences
ISSN: 2583-2727 (Online)
Volume 04| Issue 01 | Jan.-Feb. | 2024
Journal homepage: https://gjrpublication.com/gjrecs/

Review Article

A Review on Leveraging Handwritten Tulu Characters for Machine Learning in Optical Character Recognition

*Prof. Sudarshan K.[1], Dr. Sandeep Bhat[2]

[1]Research Scholar, Srinivas University, Mangaluru, Karnataka, India
[2]Professor, Department of Computer Science & Engineering, Srinivas Institute of Technology, Mangaluru, Karnataka, India

DOI: 10.5281/zenodo.10685054 **Submission Date:** 29 Feb. 2023 | **Published Date:** 21 Feb. 2024

***Corresponding author:** Prof. Sudarshan K.
Research Scholar, Srinivas University, Mangaluru, Karnataka, India

Abstract
Character identification in the Tulu language, spoken along the southwestern Indian coast, faces major obstacles because standardized datasets are not available. This review addresses this challenge by integrating handwritten Tulu characters into machine learning models, recognizing the script's complexity and unpredictability. Leveraging human-created data and related works done in other languages, this study pioneers an innovative approach to handwritten dataset production and bridging the gap in reliable recognition methods. By infusing cultural essence into the machine learning process, these handwritten samples, humanize algorithms, forging a link between script heritage and technological advancement. Through meticulous data augmentation, model training, and iterative refinement, the study aims to enhance recognition resilience and accuracy. OCR on handwritten materials is crucial, and handwritten character recognition has always been a frontier area of study in pattern recognition and image processing. Despite the fact that numerous studies have been conducted on foreign scripts such as Chinese, Japanese, Arabic characters, and even in most of handwritten Indian scripts, just a few studies have been conducted on handwritten character identification particularly in South Indian scripts like Tulu. This study gives an overview of work done in Tulu language and other languages suggest a proposed method of data set creation and conclude the need for Tulu language handwritten character data set creation and recognition. Proposed method for dataset creation is mentioned along with the architecture of the system.

Keywords: Tulu Language, Character Recognition, Handwritten Dataset, Machine Learning, Cultural Preservation, Dataset Creation.

1. INTRODUCTION
Character recognition in handwritten Tulu, an indigenous language of southwest India, poses a distinct problem. Character recognition for other Indian languages has advanced, but Tulu script has received less attention, hence there are few resources and studies in this area. A number of groundbreaking studies have established the foundation for handwritten Tulu character identification including the mapping of characters and their recognition through OCR, with MATLAB chosen for implementation[1], identification of Tulu handwritten characters from manuscripts on palm leaves [2], classification of handwritten Tulu characters using PNN and Deep CNN models[3], contrasting deep learning (Deep CNN) and shallow learning (ANN, SVM, AdaBoost) approaches for character recognition using pre-processing, feature extraction, classification, and recognition stages [4], The recognition algorithm AdaBoost is based on Haar features, and the Cascade trainer model for comparison[5] comprising 45 consonants, 13 vowels, and vowel diacritics that demonstrated the accuracy of Tulu character recognition based on recall and precision tests, a framework for various machine learning techniques such as neural network classifiers, statistical classifiers, support vector machines, and multiple classifier combinations for character recognition[6] . Few studies examined the subtleties of Tulu script, emphasizing the difficulties brought on by its unique character set, compound characters, and the existence of vowel modifiers. Some works done in related languages like Malayalam, Arabic, Kannada is also discussed here. In order to

provide the groundwork for further research in this area, some studies focused on creating a draft dataset for Tulu characters.

2. Design of a System for Recognizing Characters

Character recognition involves several crucial steps, including preprocessing, segmentation, feature extraction, and classification. These successive actions work together to extract useful data and refine image characteristics. Every stage's output flow into the next with ease, generating a dynamic process where data is gradually improved. This methodical process guarantees that the original raw input goes through a number of changes, each of which adds to a more complex comprehension of the image's characters. The cooperative character of these steps—from image correction to feature extraction—optimizes the character recognition process as a whole, producing precise and effective classification outcomes. Figure 1 shows Block Schematic for the General Character Recognition Program.

2.1 Pre processing:

The act of formatting an image before using it is known as image preprocessing for model training and inference. This includes, among other things, resizing, aligning, and color adjustments. The kind of image and the intended use determine which preprocessing method is best. Here are some techniques to improve image quality and conformance:

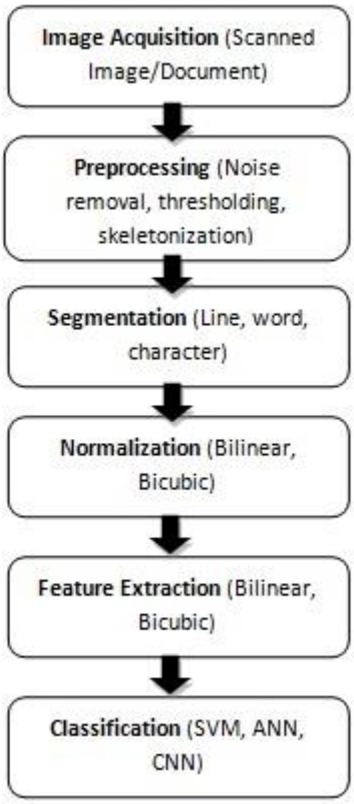
Image Acquisition (Scanned Image/Document)
Preprocessing (Noise removal, thresholding, skeletonization)
Segmentation (Line, word, character)
Normalization (Bilinear, Bicubic)
Feature Extraction (Bilinear, Bicubic)
Classification (SVM, ANN, CNN)

Figure 1: Block Diagram of General Character Recognition System

2.1.1 Noise Removal:

Any loss in image quality caused by external disturbance is referred to as noise. Pen quality, ink color, age, type of paper used, and other factors all have an impact on the quality of handwritten documents. Examples of noise include Gaussian noise and salt and pepper noise. Filtering strategies can reduce these disruptions to some degree.

2.1.2 Thresholding:

The technique of dividing the foreground (ink) from the background (paper) is called thresholding. Given a threshold T between 0 and 255, replace all pixels with grey levels less than or equal to T with black (0) and the remaining pixels with white. Certain items will be concealed and the number of objects displayed will decrease if the threshold is set too low. If it is excessively high, we might offer background information that is not desired. It is possible to apply the appropriate threshold value locally or globally.

2.1.3 Skeletonization:

A preprocessing method called "skeletonization" sharpens an image by converting binary-valued picture regions into lines that stand in for the skeletons of the regions.

2.2 Segmentation:

The stage of segmentation includes line, word, and character segmentation among others. Projections, connected component labeling, and white space and pitch are the foundations of character segmentation techniques.

2.3 Normalization:

It's the procedure of converting an arbitrary image size to a standard image size. The variance in character between classes is eliminated by this size normalization. Bilinear and Bicubic interpolation are two methods of size normalization.

2.4 Feature Extraction:

This process separates the original features into a new set by emphasizing significant patterns and preserving relevant data. The data representation gets smaller as a result. In summary, feature extraction builds upon existing features to produce new ones, while feature selection chooses specific attributes.

2.5 Classification:

Determining which model best fits the input character image is the aim of the classification stage. The two primary phases of the classification process are learning and decision-making. The system uses a training set of samples to learn about the pertinent properties of the model classes during the learning step. The next step in the decision-making process is to try and predict the model closer that corresponds to this character image.

3. Various Feature Extraction and Classification methods

Manimozhi, Manoj in their work [1] proposed the development of an efficient translation system for Tulu to Kannada using Optical Character Recognition (OCR) techniques. Using MATLAB for implementation, the framework enables character mapping and OCR recognition. The paper cites a number of studies and research papers on OCR for various Indian languages. Tulu is also being promoted in government Karnataka by the Tulu Academy, which has also released an electronic version of the Tulu script [11] [12]. Employing Optical Character Recognition (OCR) techniques, the work entails developing an effective translation system from Tulu to Kannada. The procedure is broken down into blocks on the system flow chart, and includes steps like feature extraction, template matching, and image pre-processing. The report makes reference to different approaches that have been suggested for feature matching and extraction for different Indian dialects, as well as the application of template matching for other Indian contents such as Gujarati, Dogri, and, and Gur. With an accuracy of 80.5%, the proposed work presents a portion of the 245 total tested samples, of which 197 samples were successfully recognized. Tesseract validations are also carried out. According to the work, the variety in writing styles, shapes, and directions makes it difficult to recognize handwritten Tulu characters. The limitations of Otsu thresholding and the challenges associated with Arabic letter recognition [4] [10] because of their dynamic writing style are also covered. Furthermore, the necessity of a good recognition framework for the handwritten arrangement and the difficulties in identifying different Indian contents are covered.

The goal of P J Antony and C K Savitha's proposed work [2] was to identify handwritten Tulu characters from palm leaf manuscripts. Using an automated tool that combines edge detection and thresholding techniques, the suggested approach binarises the image. Using connected component analysis and an extra projection profile, line and character segmentation is accomplished. Here, segmented Tulu characters are identified and features are effectively extracted using a deep convolution neural network (DCNN) model [13] with a recognition rate of 79.92%. The results were validated model was extended to tasks involving handwritten character recognition using the AMADI_LontarSet benchmark dataset [14]. The results showed that their method outperforms the state-of-the-art models at this time.

A For the ancient South Indian language of Tulu, C K Savitha and P J Antony proposed an offline machine learning-based handwritten character recognition (HCR) system [3]. The recommended method focuses on image segmentation using skeletonization techniques, adaptive thresholding based binarization [15], and noise reduction in conjunction with connected component analysis. The PNN and Deep CNN [16] classifiers are compared in this study. A combination of zonal and WT based 'feature extraction methods' is proposed to raise the 'accuracy' of palm leaf dataset recognition to 86.12% in PNN. Deep CNN improves classifier efficiency over state-of-the-art methods, cutting execution time to 32 seconds by combining feature extraction with a classifier model.

C. K. Savitha and P. J. Antony focused on the application of shallow and deep machine learning techniques for the offline handwritten character recognition of south Dravidian Tulu scripts in the proposed work [4]. Character recognition and classification are achieved through the extraction of zone-wise density and gradient features using shallow learning techniques like Artificial Neural Networks (ANN), Support Vector Machines (SVM), and AdaBoost, and deep learning techniques like Deep Convolution Neural Network (Deep CNN) classifier. A comparative analysis shows that Deep CNN outperforms shallow learning techniques with an efficiency of 98.49% for isolated Tulu characters from modern documents, and 80.49% for isolated characters from Tulu palms. The number and caliber of training samples utilized determine how accurate recognition is. The number and caliber of training samples utilized determine how accurate recognition is. [14]. AdaBoost functions effectively when provided with high-quality training data. If Tulu palm leaf manuscripts have damaged characters, the manuscripts' accuracy is reduced. An improved result for recognition is obtained with more samples. The system is capable of identifying characters with varying dimensions. The efficiency of the system can be increased by handling distorted characters with the creative preprocessing methods. Expanding the work to include word detection was also mentioned by the authors.

In Tulu Script Handwritten Character Recognition System based on Haar Features [5] a framework for translating Tulu script to Kannada characters and identifying it using an offline recognition mechanism was proposed by Antony P. J., Savitha C.K., and Ujwal U J [17]. The process comprised digitization, feature extraction, training, segmentation, binarization [18][19], and classification [20]. The Cascade trainer model was utilized for comparison, and the AdaBoost algorithm, which is based on Haar features, was utilized for recognition. The handwritten character samples from students, comprising 13 vowels, 45 consonants, and vowel diacritics, made up the dataset. Based on precision and recall metrics, the findings demonstrated the Tulu characters' recognition accuracy. The lack of research into handwritten character recognition was one of the limitations, and more development was required. More research is required to identify larger samples using word, sentence, and other recognition techniques as the system has only been tested on a limited number of samples.

People can gain more experience by preserving old archives that have a readable and editable structure. There are many Tulu historical documents [21] available in handwritten form, making Tulu one of the five notable Dravidian dialects. With numerous connected character combinations, Tulu scripts are intricate and full of patterns. Machine recognition will henceforth be a significant difficulty. The framework put forth by C. K. Savitha[6] listed the methods used for handwritten character recognition and highlighted the key characteristics of Tulu script. The work discusses the need for machine recognition of handwritten documents, particularly the Tulu script [22] and the challenges involved in this task. It highlights the historical significance of the Tulu script and the importance of recognizing and understanding it for research purposes. Proposed work references various machine learning techniques such as neural network classifiers, statistical classifiers, support vector machines, and multiple classifier [23] combinations for character recognition. It is also mentioned in the work that the recognition system is trained and tested using handwritten character samples in Malayalam [24][25] and Tulu scripts.

Kannada handwritten characters and numerals on paper using the OCR system [7] an efficient feature extraction and classification technique is the goal of Saleem Pasha, M C Padma's proposed method for handwritten Kannada characters and numerals recognition. To transform input images into a format that is appropriate for feature extraction, preprocessing methods are used. Recognization is accomplished by an artificial neural network classifier, which is trained using wavelet transform and structural features [26]. 91.00% for characters and 97.60% for numerals is the average accuracy achieved by the method. There are 1000 handwritten Kannada numerals and 4800 handwritten Kannada characters in the dataset. The research comes to the conclusion that the suggested approach has the potential to handle more complicated characters and numerals in the future and demonstrates encouraging results when it comes to handwritten Kannada.

Few studies have addressed Persian scripts despite the wide range of applications for digit, letter, and word recognition. Persian Handwritten Digit, Character, and Word Recognition is a proposed work by Mahdi Bonyani1 et al. Deep neural networks are used with various DenseNt and Xception architectures using Deep Learning [8]. Data augmentation and test time augmentation are used to further enhance the effectiveness of the networks. Using machine learning algorithms, the paper focuses on optical character recognition (OCR) for handwritten Persian text. It recognizes numbers, letters, and words using the DenseNet121, DenseNet161, DenseNet169, Xception, and DenseNet201 architectures. Training and testing are conducted using the HODA and Sadri databases. The suggested approach outperforms earlier research, achieving recognition rates of 99.49% [28] for digits and 98.10% [27] for words. One of the constraints is the small number of previous studies on handwritten Persian words.

A method that involved using statistical analysis [29] and feature extraction techniques to develop an optical character recognition (OCR) system for Telugu was proposed by N Prameela and P Anjusha [9]. QDA and SVM, or support vector machine and discriminate classifier, were the ML techniques that were employed. There were two hundred samples in all from the Telugu character database's forty-seven categories. 80.6% and 87.6%, respectively, of the SVM and QDA classifiers' recognition rates were shown in the results. In addition to stressing the necessity of shape- and font-dependent pre-processing and feature extraction, the conclusion underlined the viability of the suggested OCR system for Telugu character recognition. It is evident that the two character shapes are similar to one another, with the exception of the areas at the top and bottom. To sum up, the necessity for a workable OCR system for Telugu character recognition drove the design, methodology, and execution.

With the goal of being able to convert printed text or poetry in Kannada script without any vocabulary restrictions, Lipi Gnani [10] is a Kannada OCR that was created from the ground up. The suggested solution performed exceptionally well at identifying old, printed pages in Kannada. Numerous elements, including Hale-gannada characters, punctuation marks, and numerals in both Indo-Arabic and Kannada, are recognized by the system. Machine learning techniques were used in the development of the OCR system, such as deep neural networks with The hidden layers' LSTM cells and connectionist temporal classification in the output layer. Languages like Kannada, Konkani, Tulu, and Sanskrit were all printed using Kannada script, and annotated ground truth was included in a benchmarking test dataset. With respect to current OCR systems, Lipi Gnani OCR's results on these datasets demonstrated notable advancements. Results in Kannada, Sanskrit, Konkani, and Tulu datasets showed that Lipi Gnani's word level recognition accuracy was superior to Google's Tesseract OCR. One of the system's drawbacks is that it only uses the recognition process to produce results; it does not have a dictionary-based post processing. The authors' 5000 printed pages in multiple languages that comprised the dataset are not publicly accessible. The system performed exceptionally well on multiple datasets, surpassing Google's Tesseract OCR.

4. Proposed Method for Dataset Creation

The proposed method involves engaging 100 and more writers to handwrite the complete set of Tulu characters in a 9x6 grid on paper. This initiative aims to create a diverse dataset reflecting various writing styles and nuances. Writers will follow specific guidelines to maintain consistency in character size, orientation, and placement within the grid. The sample sheet of paper where one writer has written the characters is as shown in figure 2.

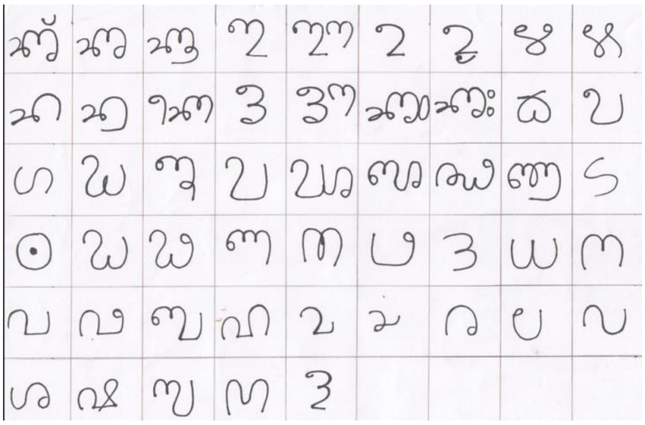

Figure 2: Sample Data Collected from One Writer

Following collection, high-resolution scanning or imaging methods will be used to transform the handwritten grids into a digital format. The characters' accuracy and clarity will be preserved through this digitization process. After digitization, preprocessing operations such as segmentation, normalization, and noise reduction will be applied to the dataset. By isolating each character, these steps will get them ready for further classification. The characters will be thoroughly annotated, matching each character to its corresponding identity in Tulu. The basis for training deep learning or machine learning models—such as CNNs or RNNs—for character recognition will be this annotated dataset. Through the use of validation techniques, the performance of the trained model will be assessed, allowing for iterative improvements to improve accuracy and applicability. A dependable Tulu character dataset for the creation of accurate character recognition algorithms is the ultimate objective of this methodology.

5. CONCLUSION

This review work covers optical character recognition systems especially for handwritten Malayalam, Tulu, Persian and Kannada scripts. In addition, different segmentation strategies and classifiers with various features are investigated. Despite its significance and urgency, this topic has not received enough attention from academics. One of the main issues in this field is the lack of a benchmark database for handwritten characters in the majority of languages for testing research findings. So there is a need to create a standardized dataset and make it available to the researchers for doing further research especially in Tulu language. In an effort to build a comprehensive and diverse dataset, 100 different writers have contributed handwritten Tulu characters in a 9x6 grid. This collection will provide as a fundamental resource for training strong character recognition models through careful augmentation for creating more datasets, digitization, preprocessing, and annotation. The potential to develop accurate and versatile Tulu character recognition systems can be realized by utilizing machine learning or deep learning techniques with this dataset. Tulu script recognition technology is expected to advance significantly with the comprehensive dataset that is produced by the method's systematic approach, which guarantees the preservation of writing style nuances. Accurate Tulu script recognition has significant implications for future technological developments. The recognition system has a wide range of uses, from language preservation programs to document digitization. A few possible uses for this technological advance include digitizing old Tulu manuscripts, automating data entry procedures, and facilitating effective information retrieval. The synergy of ML or deep learning techniques with a comprehensive and systematically generated dataset holds the key to unlocking the full potential of Tulu character recognition systems.

References

1. Dr..Manimozhi and Dr. Manoj challa "An Efficient Translation of Tulu to Kannada South Indian Scripts using Optical Character Recognition" in the Proccedings of the Fifth lnternational Conference om Computing Methodologics and Communication (1CCMC 2021).
2. C K Savitha and P J Antony, "Segmentation and recognition of characters on Tulu palm leaf manuscripts in International Journal of Computational Vision and Robotics, 2019 Vol.9 No.5, pp.438 – 457.
3. CK Savitha and P J Antony "PNN and Deep Learning Based Character Recognition System for Tulu Manuscripts" in the International Journal of Engineering and Advanced Technology (JEAT) ISSN: 2249-8958, Volume-8 Issue-5, June 2019.
4. C K Savitha and P J Antony "Machine Learning Approaches for recognition of offline Tulu Handwritten Scripts" in Journal of Physics Conference Series 2018.
5. Antony P. J. Savitha C.K., Ujwal U J "Haar Features based Handwritten Character Recognition System for Tulu Script" IEEE International Conference on Recent Trends in Electronics Information Communication Technology, May 20-21, 2016, India 978-1-5090-0774-5/16/$31.00 © 2016 IEEE 6.
6. C K Savitha and P J Antony, "A framework for recognition of handwritten South Dravidian Tulu script" in Conference on Advances in Signal Processing (CASP) 2016.
7. Saleem Pasha and MC Padma "Handwritten Kannada Character Recognition using Wavelet Transform and Structural Features" in International Conference on Emerging Research in Electronics, Computer Science and Technology – 2015.
8. Mehdi Bonyani, Simindokht Jahangard and Morteza Daneshmand "Persian Handwritten Digit, Character and Word Recognition Using Deep Learning" in International Journal on Document Analysis and Recognition (IJDAR) 2021.
9. N Prameela P Anjusha R Karthik "Off-line Telugu Handwritten Characters Recognition using optical character recognition" in International Conference on Electronics, Communication and Aerospace Technology ICECA 2017.
10. H. R. Shiva kumar & A. G. Ramakrishnan Lipi Gnani "A Versatile OCR for Documents in any Language Printed in Kannada Script" in ACM Transactions on Asian and Low-Resource Language Information Processing, Vol. 1, No. 1, Article 1. Publication date: February 2020.
11. Roshan Fernandes, Anisha P Rodrigues "Kannada Hand written script Recognition using machine learning techniques "in 2019 IEEE International Conference on distributing computing, VLSI, electrical Circuits and robotics (DISCOVER).
12. S.Kowsalya ,P.S Perisamy " Recognition of Tamil hand written character using modified neural network with aid of elephant herding optimization" Springer 2019.
13. Liu, C., Liu, J., Yu, F., Huang, Y. and Chen, J. (2013) 'Handwritten character recognition with sequential convolutional neural network', in 2013 International Conference on Machine Learning and Cybernetics (ICMLC), IEEE, Vol. 1, pp.291–296.
14. Burie, J-C., Coustaty, M., Hadi, S., Kesiman, M.W.A., Ogier, J-M., Paulus, E., Sok, K., Sunarya, I.M.G. and Valy, D. (2016) 'ICFHR2016 competition on the analysis of handwritten text in images of Balinese palm leaf manuscripts', in 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), IEEE, pp.596–601.
15. Ingle, Prashant Devidas, and Parminder Kaur. "Adaptive thresholding to robust image binarization for degraded document images." In Intelligent Systems and Information Management (ICISIM), 2017 1st International Conference on, IEEE, pp. 189-193, 2017.
16. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
17. Poojary Jayakar D Tulu Lipi Terile 2012 A Book on Tulu Script.
18. Chacko, Anitha Mary MO, P. M. Dhanya. "Combining Classifiers for Offline Malayalam Character Recognition." In Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2, pp. 19-26. Springer International Publishing, 2015.
19. Shanjana, C., and Ashish James. "Character segmentation in Malayalam Handwritten documents." In Advances in Engineering and Technology Research (ICAETR), 2014 International Conference on, pp. 1-4. IEEE, 2014.

20. Viola, Paul, and Michael J. Jones. "Robust real-time face detection." International journal of computer vision 57, no. 2 (2004): 137-154.
21. Jayakar D Poojary, “Tulu Lipi Terile”, A Book on Tulu Script., 2012 Publisher-J.P. Prakashana Mumbai.
22. Fischer, Andreas, "Handwriting recognition in historical documents", PhD diss., 2012.
23. Liu, Cheng-Lin, Hiromichi Fujisawa. "Classification and learning for character recognition: comparison of methods and remaining problems." InInt. Workshop on Neural Networks and Learning in Document Analysis and Recognition. 2005
24. Jayakar D Poojary, “Tulu Lipi Terile”, A Book on Tulu Script., 2012 Publisher-J.P. Prakashana Mumbai.
25. K. Padmanabha Kekunnaya, “A comparative study of Tulu dialects”, 1994. Publisher –Rashtrakavi Govinda Pai Research Centre Udupi.
26. B.V. Dhandra, Shashikala Parameshwarapa and Gururaj Mukarambi, “Kannada Handwritten Vowels Recognition based on Normalized Chain Code and Wavelet Filters”, International Journal of Computer Applications (0975 – 8887), Recent Advances in Information Technology, pp. 21–24, 2014.
27. Sadri, J., Yeganehzad, M.R., Saghi, J.: A novel comprehensive database for offline Persian handwriting recognition. Pattern Recognit. 60, 378–393 (2016).
28. Sarvaramini, F., Nasrollahzadeh, A., Soryani, M.: Persian handwritten character recognition using convolutional neural network. In: Iranian Conference on Electrical Engineering (ICEE), pp. 1676–1680 (2018).
29. C.Vikram and C.Shoba Bindhu,”Hand written character Recognition for Telugu Script using Multilayer Perceptrons”,IJARCET-VOL2

CITATION

Sudarshan K., & Sandeep B. (2024). A Review on Leveraging Handwritten Tulu Characters for Machine Learning in Optical Character Recognition. In Global Journal of Research in Engineering & Computer Sciences (Vol. 4, Number 1, pp. 67–73). https://doi.org/10.5281/zenodo.10685054