DOI: http://dx.doi.org/10.26483/ijarcs.v16i1.7182

ISSN No. 0976-5697

Volume 16, No. 1, January-February 2025

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

BUILT A DATASET OF GUJARATI ISOLATED HANDWRITTEN CHARACTERS AND RECOGNITION THROUGH DEEP LEARNING

Jitendrakumar B. Upadhyay
Shrimad Rajchandra Institute of Management and Computer Application,
Uka Tarsadia University, Bardoli, Gujarat, India.

Jitendra Nasriwala
Babu Madhav Institute of Information Technology,
Uka Tarsadia University,
Bardoli, Gujarat, India.

Abstract: In the current era with the rise of new machine learning algorithms, particularly deep learning, the demand for large, high-quality datasets has grown significantly, especially in handwritten character recognition (HCR). While several Indian languages have publicly available benchmark datasets, a few, including Gujarati, still lack such resources. This paper addresses an attempt to build a dataset for Gujarati isolated handwritten characters and to recognize the isolated Gujarati handwritten vowels and consonants. The dataset is collected from 692 writers of varying ages, genders, qualifications, and professions. The dataset consists of 63,664 samples for 46 classes including 34 consonants and 12 vowels where 1384 images of each character. The proposed model was run with an 80:20 training and testing ratio, using 7, 10, 20, 30, & 40 epochs. The model showed promising results and achieved the highest training accuracy 90.92%, and the highest testing accuracy 89.51%.

Keywords: Handwritten Character Recognition, Gujarati, Deep Learning, Convolution Neural Network

I. INTRODUCTION

Handwritten Character Recognition (HCR) has emerged as one of the most captivating and challenging areas of research in image processing and pattern recognition in recent years. While recognizing handwritten characters is a simple task for the human mind, it poses significant challenges and complexities for machines. In recent years, various deep learning, machine learning models, and feature extraction methods have been applied using the benchmark datasets to achieve good accuracy rates. A well-constructed dataset ensures diversity, capturing a wide range of handwriting styles, sizes, and orientations, which is essential for generalization.

Benchmark datasets for handwritten characters are widely available across well-known languages, serving as crucial resources for advancing character recognition. Among the languages, the English language has prominent datasets like the Modified National Institute of Standards and Technology (MNIST), National Institute of Standards and Technology (NIST) , Center of Excellence for Document Analysis and Recognition (CEDAR), and Centre for Pattern Recognition and Machine Intelligence (CENPARMI) [1,2]. Similarly, datasets exist for other major scripts, including Arabic, Chinese, Japanese, Korean, etc. [3]. In the case of Indic scripts, a few datasets are available publicly, such as Devanagari [4,5], Bangla (e.g., CMATERDB, ISI Kolkata, and BanglaLekha-Isolated [5,6]), Telugu [7], Tamil [2], Gurmukhi [3], etc. These datasets provide essential benchmarks for handwritten character recognition across different languages and scripts. Despite advances in HCR, there remains a significant gap in the availability of comprehensive datasets for Gujarati languages [3]. The "Technology Development for Indian Languages (TDIL)" provides a dataset for Gujarati, which includes around 30 characters and 12,860 sample images [8,9]. However, due to limitation such as lacking of a benchmark dataset of Gujarati handwritten characters, researchers often rely on their datasets and focus on a subset of characters. Motivated by this gap, we developed a benchmark dataset for handwritten Gujarati isolated characters, encompassing both vowels and consonants, which will be made publicly available for researchers.

Deep Learning is a subset of machine learning [10]. Deep Learning has plenty of tools for pattern recognition. With the help of intrinsic tools, it is easier to pinpoint significant solutions for image processing, speech recognition, and NLP problems. Deep Learning is specially used for character recognition and segmentation from the image. To achieve a good recognition rate extracted features play a vital role. But to extract manually features from images requires more time and it is a very tedious job and they cannot process raw photos either. On the other side, there is no need to describe any explicit features for Deep Convolutional Neural Networks (CNN) instead they make use of raw pixel information to generate the best features for classification [11]. Due to the advantages of deep learning techniques, developed a customized CNN and implemented it on a self-generated dataset.

The rest of the paper is as follows: Section 2 reviews prior research on the dataset-generation process for HCR and deep learning. Section 3 introduced about Gujarati Language Script. Section 4 introduces the proposed approach to the dataset generation process for the Gujarati language Handwritten Character and Character Recognition Process through deep learning. Section 5 discussed a detailed analysis of the dataset, character recognition, and outcomes. Finally, Section 6 presents the conclusion.

II. RELATED WORK

This section focused on the examination of the methodologies and challenges involved in constructing HCR datasets, emphasizing the diverse approaches and technologies utilized in this field and Handwritten Character recognition where the language is Gujarati and the methodology uses a deep learning approach.

Hebbi *et al.* [2] developed a dataset for handwritten Kannada characters, comprising 1,30,981 samples across 85 characters, including vowels, consonants, modifiers, Ottaksharas, and Yogavahagalu. Data were collected from 500 writers of varying ages and genders. They applied erosion, dilation, and noise removal using median blur. Biswas *et al.* [4,6] presented a methodology for constructing a dataset for handwritten Bangla characters, which consists of 1.66,105 samples across 84 classes collected from 2,000 writers. This dataset comprised 50 Bangla basic characters, 10 Bangla numerals, and 24 selected compound characters. They employed preprocessing techniques including binary inversion, noise removal, and edge thickening. Dongre *et al.* [1] described the process for creating a dataset of Devanagari handwritten characters, gathering data from 750 individuals including students of schools and colleges, office staff, workers, housewives, and senior citizens. This dataset includes 5,137 numerals and 20,305 characters.

A dataset for handwritten Telugu characters was presented by VaraLakshmi *et al.* [7], with data taken the 100 different sets of characters, each sample consists of 983 Telugu characters. While some researchers have utilized the datasets available from the Technology Development for Indian Languages (TDIL) for Gujarati handwritten characters [8,9], this dataset contains 12,840 image samples of 30 characters. Consequently, many researchers have collected or generated their datasets for recognition purposes through applied preprocessing, segmentation, and labeling steps. Among these, Thaker *et al.* [12] present different algorithms for preprocessing including grayscale conversion, noise removal, contrast adjustment, binarization, thinning, and segregating handwritten datasheets containing 30 isolated characters written in Gujarati script. Sharma *et al.* [13] generated a total of 88,000 isolated character images, divided into 44 classes, which include six vowels, one numeral, and consonants, with a few classes combining characters, such as the character ‘Ga,’ which merges two classes. This dataset was collected from 2,000 individuals across different age groups, educational backgrounds, and professions.

Pareekh *et al.* [14] describe a supervised classifier approach based on CNN and ML for the recognition of handwritten Gujarati characters. They created 10,000 images which were collected from 250 different people and segmented individual characters of size 28 X 28 pixels. A success rate of 97.21% is obtained using CNN and 64.48% using MLP. Rajygore *et al.* [15] collected 58,000 Gujarati handwritten character images from individuals aged 15 to 70 years, resulting in a dataset comprising 74 classes, including consonants, vowels, connected characters, special characters, and numerals. The author used the LSTM classifier for recognition using Keras Library and achieved the prediction correct result from all test images is 97.00% around. Suthar et al. [16] used 83 writer Gujarati script datasets and applied pre-processing and segmentation processes to it. They used a classifier for the recognition of Gujarati Characters KNN, SVM, NB, and CNN models on their own generated datasets the accuracy is 91.1%, 97.4%, 97.2%, and 98.2% respectively. Shukla and Desai experimented using a deep learning approach to recognize handwritten Gujarati characters and numerals. They obtained 82.15% test accuracy [17]. Limbachiya et al. [18] utilized fusion mechanism such as transfer learning using five pre-trained networks including VGG16, InceptionV3, DenseNet, Nasnet, and MobileNet on handwritten Gujarati alphanumeric characters’ dataset covering 54 Gujarati character classes including 34 consonants, 12 vowels, and 10 numerals. They obtained the highest accuracy of 97% using the MobileNet pre-trained network.

Based on the literature review, identify the need to create a benchmark dataset of isolated Gujarati Handwritten Characters to implement the HCR process and get a good accuracy rate using deep learning techniques.

III. GUJARATI SCRIPT

The Devanagari script, primarily used for Hindi and Sanskrit, originated from the ancient Brahmi script and evolved through the Kutil and Nagri scripts around the 8th and 9th centuries. Influenced by a few languages such as Farsi, Turkish, Arabic, English, and Dravidian languages, Devanagari became the foundation for modern scripts like Hindi, Marathi, and Gujarati. Gujarati is a language with a rich history spanning over a thousand years, originating from the Indo-Aryan language family that evolved from Sanskrit. Spoken by more than 62 million people [18], primarily in the Indian state of Gujarat, it holds the distinction of being the 26th most widely spoken language globally.

The Gujarati script is characterized by its 12 vowels (Swar), 34 consonants, and 10-digit symbols. Additionally, it features 14 diacritic marks (Matra) corresponding to each vowel, which attach to consonants to modify their phonetic output. The vowels are shown in Fig. 1 and the consonants are shown in Fig. 2.

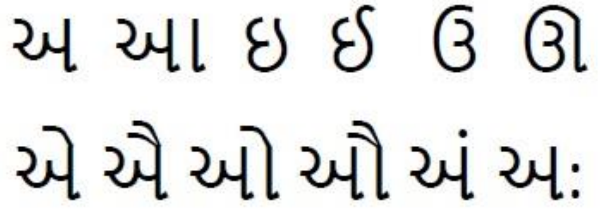


Figure 1 Gujarati Vowels

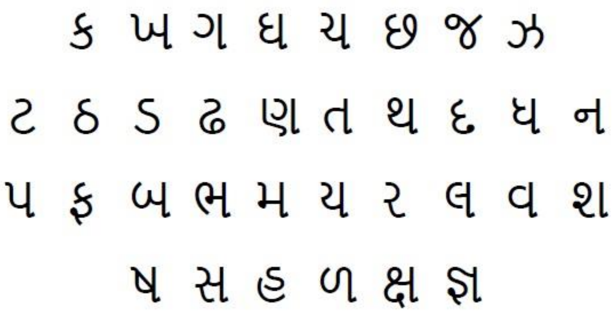


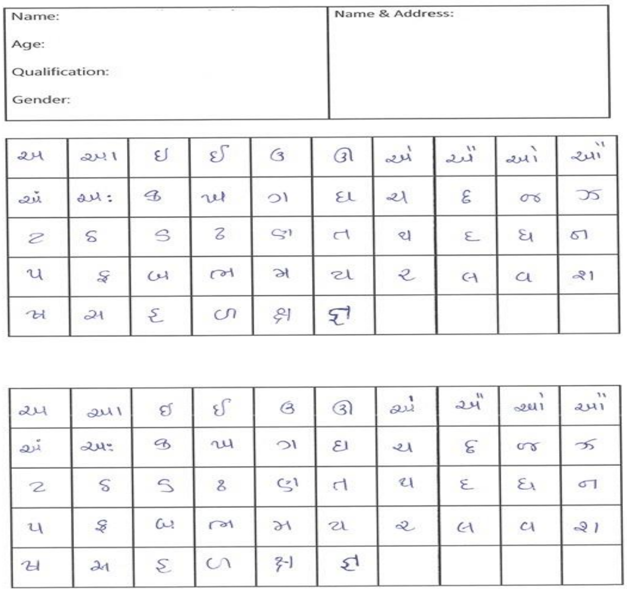
Figure 2 Gujarati Consonants

IV. METHODOLOGY

This section is divided into two sessions; The first is a dataset generation process and the second is character recognition through CNN.

4.1. Dataset Generation Process

The dataset generation process begins with the creation of a sample datasheet that guides writers on how to accurately record the data. The datasheet is designed to be easily understandable, ensuring that writers from various domains and with different working styles can follow the proper format. So, the initial stage of data generation is followed by data collection, where samples are gathered from a diverse group of writers, each contributing their unique handwriting styles. Each writer received an A4-sized sheet shown in Fig. 3, meticulously divided into a grid, providing a structured framework for writing Gujarati characters. This approach captures a broad spectrum of handwriting variations, reflecting differences in writing patterns across age groups, backgrounds, and individual characteristics. Once the raw data is collected, it enters a preprocessing stage to prepare the images for further analysis. Preprocessing involves several critical steps, including converting images to grayscale, applying thresholding to enhance contrast, and removing noise [4,7,12,13,16]. This stage is vital for standardizing the dataset and ensuring high image quality, as it simplifies the data's complexities while preserving essential features. Additionally, image resizing is applied to maintain uniformity across the dataset. After preprocessing, segmentation techniques are employed to isolate individual characters from the images. This process utilizes methods such as contour detection to identify segment boundaries and the Hough Line Transform technique for enhanced accuracy [14].



Name: Age: Qualification: Gender:	Name & Address:

અ	આ	ઇ	ઈ	ઉ	ઊ	એ	ઐ	ઓ	ઔ
અં	અઃ	ક	ખ	ગ	ઘ	ચ	છ	જ	ઝ
ટ	ઠ	ડ	ઢ	ણ	ત	થ	દ	ધ	ન
પ	ફ	બ	ભ	મ	ય	ર	લ	વ	શ
ષ	સ	હ	ળ	ક્ષ	જ્ઞ				

અ	આ	ઇ	ઈ	ઉ	ઊ	એ	ઐ	ઓ	ઔ
અં	અઃ	ક	ખ	ગ	ઘ	ચ	છ	જ	ઝ
ટ	ઠ	ડ	ઢ	ણ	ત	થ	દ	ધ	ન
પ	ફ	બ	ભ	મ	ય	ર	લ	વ	શ
ષ	સ	હ	ળ	ક્ષ	જ્ઞ				

Figure 3 Gujarati isolated handwritten character datasheet sample

Once the contours are identified, they are filtered based on their size to eliminate irrelevant or smaller elements that are not part of the table structure. In real-world images, noise and other small features can interfere with contour detection, leading to false positives. By imposing size constraints, only contours that match the dimensions of typical table cells are retained, ensuring the accuracy of the segmentation process. The contours are then organized in a specific order typically from top to bottom and left to right so that the table cells can be processed in the same sequence as they appear in the table. Finally, the extracted cells, once isolated are stored in respective directories which are the labeling 0 to 45 having a sequence of characters 'અ' , 'આ' , ... , 'જ્ઞ', as per their cell sequence number in the table resulting in the generation of a dataset of isolated Characters. Fig. 4 shows the dataset generation process.

Through the dataset generation process, a total of 692 writer's datasheets were collected with different backgrounds and generated a total number of 63,664 images of 46 characters including 12 vowels and 34 consonants where each character having 1384 images. As per the our best of knowledge, these are large amounts of datasets of pure isolated handwritten character sets in Gujarati language.

4.2. Character Recognition

The generalized CNN architecture comprises a hierarchical structure that processes images by applying a sequence of convolution layers, pooling layers, fully connected layers, and flattened layers. In this paper, we proposed isolated Gujarati Handwritten Character Recognition using the customized CNN model shown in Fig. 5 to classify Gujarati isolated handwritten characters.

The architecture begins with a Conv2D layer with 64 filters and a kernel size of (7, 7), followed by batch normalization to stabilize training and a MaxPooling2D layer with a pool size of (4, 4). This combination helps capture spatial hierarchies in the input data while reducing spatial dimensions. This first Conv2D layer extracts low-level features such as edges and textures from the input images. The large kernel size helps capture more spatial context. The Conv2D layer has 3,200 trainable parameters. The second Conv2D layer was applied with 128 filters and a kernel size of (5, 5), followed by the batch normalization and followed by a MaxPooling2D layer with a pool size of (2, 2) incorporating L2 regularization to reduce overfitting. This second Conv2D layer extracts mid-level features such as corners and shapes, refining the initial representations learned by the first layer. The second Conv2D layer has 2,04,928 trainable parameters. The third Conv2D layer consists of 256 filters and a kernel size of (3, 3), batch normalization, and MaxPooling2D layer with a pool size of (2, 2). This layer captures high-level abstract features, such as complex patterns and structure, enhancing the model's ability. The third Conv2D layer has 2,95,168 trainable parameters. A Flatten layer is employed to flatten the 3-D production of the feature maps into a 1D array, passing through a fully connected (Dense) layer with 256 neurons and Rectified Linear Unit (ReLU) activation, followed by an optional dropout layer with 0.2 rate to prevent overfitting. Another Dense layer with 128 neurons applies further transformations before the final output layer, which has 46 neurons with softmax activation, making it suitable for a 46-class classification task. The model is compiled with the Adam optimizer and sparse categorical cross-entropy loss for training.

V. RESULT AND DISCUSSION

This section shows experimental analysis which was carried out using a customized CNN model on the Gujarati Handwritten Characters dataset.

5.1. Experimental setup

The proposed model was implemented in Python, using Tensorflow and Keras libraries [19]. Experiments were performed on a machine with a ASUS ROG Strix G13CHR 2024, 20 Crore, Intel® Core™ i7-14700F 14th Gen, Desktop (32GB/1TB SSD/8GB NVIDIA GeForce RTX 4060 Graphics/Windows 11), G13CHR-71470F004WS. The proposed model has been experimented with using the dataset. It utilized handwritten Gujarati characters datasets containing 63,664 images across 46 classes. From the datasets, for training purposes, utilized 80% of the dataset and the remaining 20% was used for testing purposes. The proposed model is implemented with batch size 3, and the learning rate is 0.001. The implementation was carried out using different epochs like 7, 10, 20, 30 and 40.

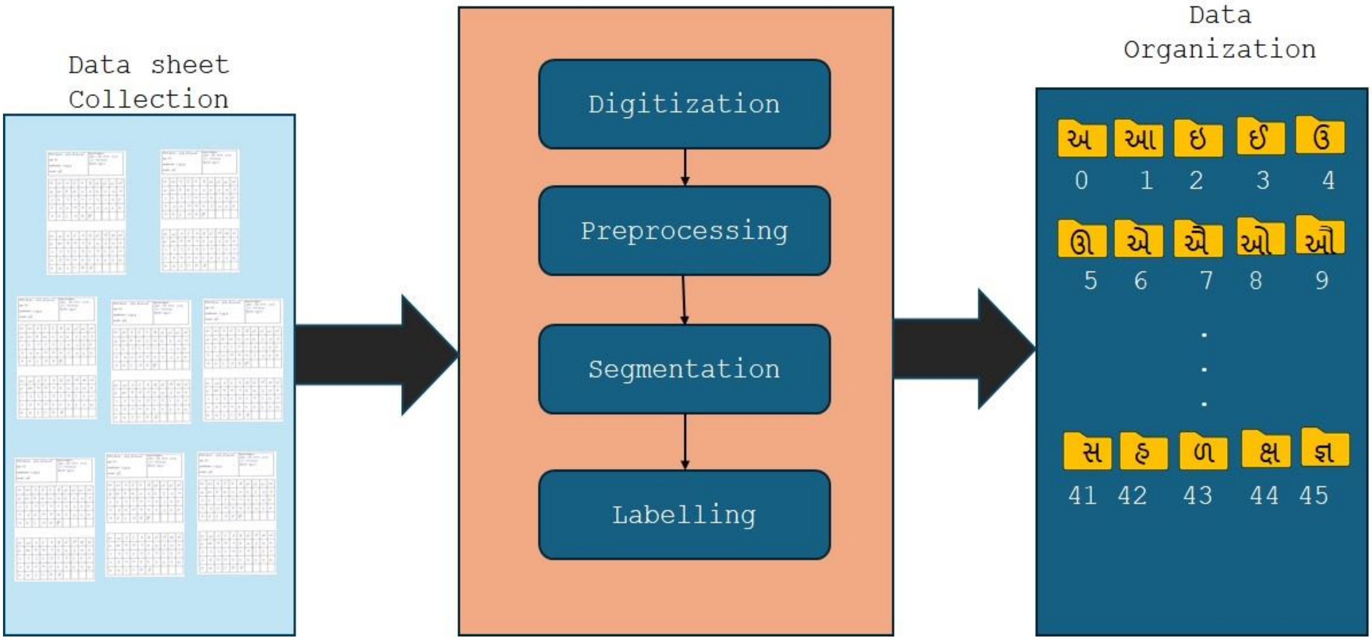


Figure 4 Gujarati isolated handwritten character dataset generation process

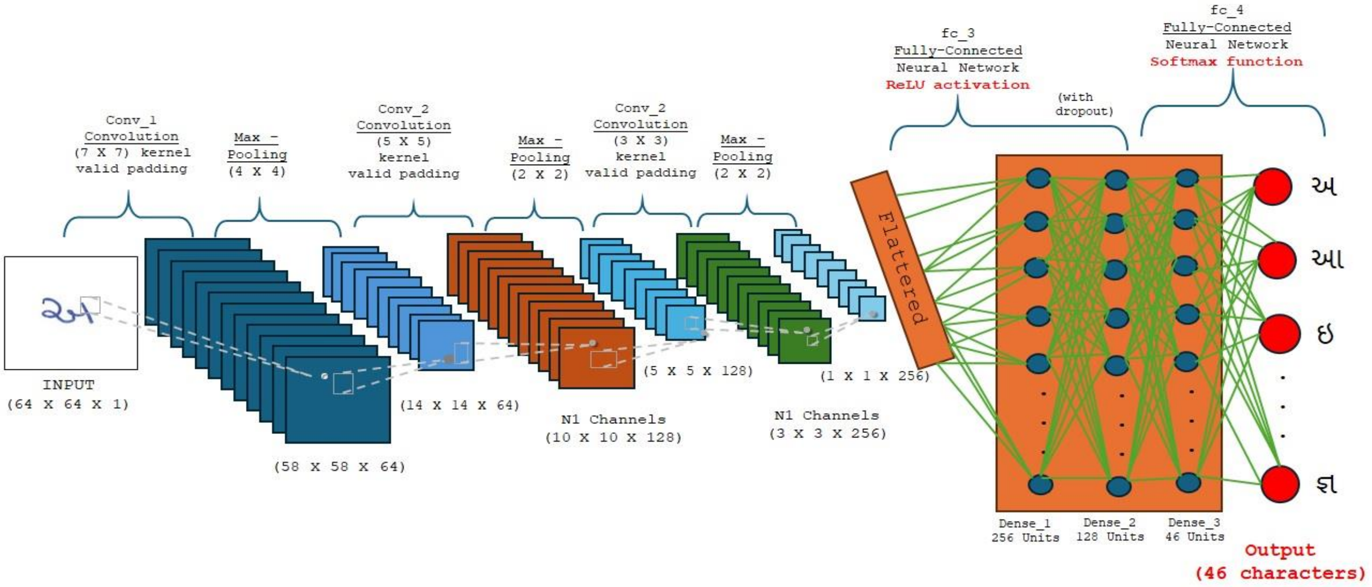

Conv_1 Convolution (7 X 7) kernel valid padding
Max - Pooling (4 X 4)
Conv_2 Convolution (5 X 5) kernel valid padding
Max - Pooling (2 X 2)
Conv_2 Convolution (3 X 3) kernel valid padding
Max - Pooling (2 X 2)
fc_3 Fully-Connected Neural Network ReLU activation
(with dropout)
fc_4 Fully-Connected Neural Network Softmax function
INPUT (64 X 64 X 1)
(58 X 58 X 64)
(14 X 14 X 64)
N1 Channels (10 X 10 X 128)
(5 X 5 X 128)
N1 Channels (3 X 3 X 256)
(1 X 1 X 256)
Flattered
Dense_1 256 Units
Dense_2 128 Units
Dense_3 46 Units
અ
આ
ઇ
જ્ઞ
Output (46 characters)

Figure 5 Customized CNN architecture for Gujarati isolated handwritten character recognition

5.2. Result discussion

The results obtained over the test dataset are evaluated through accuracy, precision, recall, and f1-score metrics, as well as a confusion matrix, which was developed for a better understanding of class-wise prediction.

Deep learning requires a large amount of data although the proposed method has provided remarkable results in different implementation phases. In the experiment achieved training accuracy of 84.48%, 86.56%, 89.10%, 90.79%, 90.92% and testing accuracy of 86.76%, 86.89%, 88.61%, 89.04%, 89.51% for epochs 7, 10, 20, 30 and 40 respectively shown in Table 1. The results found that the model is showing a good performance, still showing very few overfitting issues due to the overlapping of training and testing accuracy shown in Fig. 6, 7, 8, 9, and 10 which represent epochs vs accuracy and epochs vs loss graph for the epochs 7, 10, 20, 30 and 40. Table 1 shows the evaluation measurements like precision, recall, and F1 Score for all epochs.

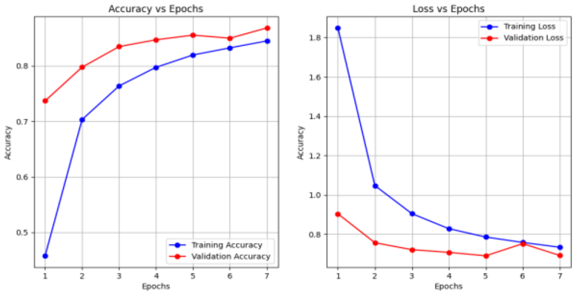


Figure 6 Accuracy Vs Epochs and Loss vs Epochs graphs for Epochs 7.

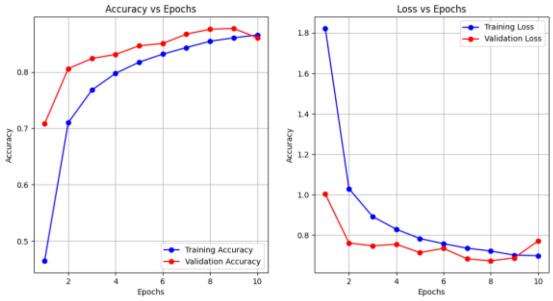


Figure 7 Accuracy Vs Epochs and Loss vs Epochs graphs for Epochs 10.

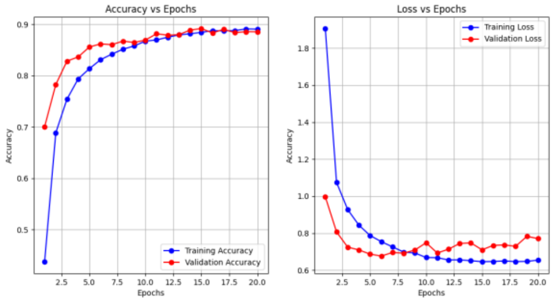


Figure 8 Accuracy Vs Epochs and Loss vs Epochs graphs for Epochs 20.

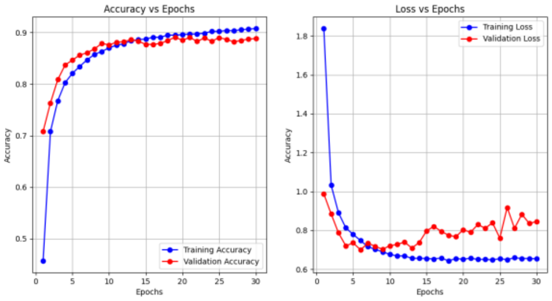


Figure 9 Accuracy Vs Epochs and Loss vs Epochs graphs for Epochs 30.

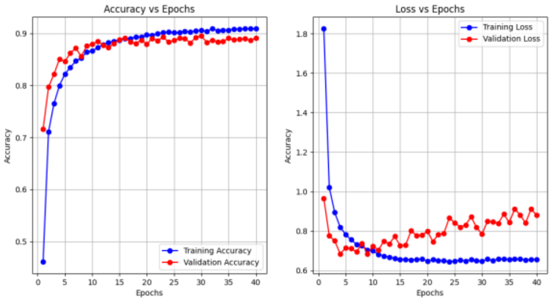


Figure 10 Accuracy Vs Epochs and Loss vs Epochs graphs for Epochs 40.

Table 1 Performance of the customized CNN model on the Gujarati Isolated Handwritten Character dataset

No of Epochs	Training Accuracy	Testing Accuracy	Precision	Recall	F1-Score
7	84.48	86.76	0.87	0.87	0.87
10	86.56	86.89	0.87	0.86	0.86
20	89.10	88.61	0.89	0.89	0.89
30	90.79	89.04	0.89	0.89	0.89
40	90.92	89.51	0.89	0.89	0.89

VI. CONCLUSION AND FUTURE WORK

Due to lacking a benchmark dataset of Gujarati Handwritten Characters, collected offline filled datasheets and generated real isolated handwritten character datasets. The result shows that 63,664 images consisting of 1384 images of each character which is the large-scale generation of original, isolated handwritten character images in the Gujarati language. We created it through real writers without the use of artificial data. The data images are stored in JPEG format for efficient storage and computational needs. This dataset will be available upon request from the authorized author and will continue to evolve as further refinements are made and new data is collected, providing a valuable resource for ongoing research and development. The proposed CNN architecture shows improvement in performance as epochs for the Gujarati isolated handwritten character recognition increase. The highest training accuracy is 90.92% and the highest testing accuracy is 89.51% obtained by the model. Although the dropout parameter was utilized in the model to prevent the overfitting issue, still few overfitting issues were found which will point out to the dataset augmentation process. The proposed isolated handwritten character recognition (HCR) model demonstrates high accuracy, with most characters achieving over 90% recognition accuracy, indicating its strength in distinguishing Gujarati isolated handwritten characters effectively.

Acknowledgment
The authors thank Dr. Rajamouli Boddula, Sanath Ezhuthaachan, and Dr. Mitesh Dwivedi for their continuous support throughout this work. They also thank the B U Tarsadia Learning Resource Centre and its dedicated staff for providing such an inspiring space and invaluable resources that made this work possible. we would also like to express our sincere appreciation to the writers who served as vital sources of information and inspiration. Additionally, thanks go to all the staff involved in the collection and data management process.

REFERENCE

[1] Dongre VJ, Mankar VH. Development of comprehensive devnagari numeral and character database for offline handwritten character recognition. Applied Computational Intelligence and Soft Computing. 2012;2012(1):871834. https://doi.org/10.1155/2012/871834

[2] Hebbi C, Mamatha HR. Comprehensive dataset building and recognition of isolated handwritten kannada characters using machine learning models. In Artificial Intelligence and Applications. 2023;1(3); pp. 179-190. https://doi.org/10.47852/bonviewaia3202624

[3] Suthar SB, Thakkar AR. Dataset Generation for Gujarati Language Using Handwritten Character Images. Wireless Personal Communications. 2024;136(4):2163-84. https://doi.org/10.1007/s11277-024-11369-9

[4] Biswas M, Islam R, Shom GK, Shopon M, Mohammed N, Momen S, Abedin A. Banglalekha-isolated: A multi-purpose comprehensive dataset of handwritten bangla isolated characters. Data in brief. 2017;12:103-7. https://doi.org/10.1016/j.dib.2017.03.035

[5] Khandokar I, Hasan M, Ernawan F, Islam S, Kabir M N. Handwritten character recognition using convolutional neural network. In Journal of Physics: Conference Series; 2021;1918(4); p. 042152. https://doi.org/10.1088/1742-6596/1918/4/042152

[6] Biswas M, Islam R, Shom GK, Shopon M, Mohammed N, Momen S, Abedin MA. Banglalekha-isolated: A comprehensive bangla handwritten character dataset. arXiv preprint arXiv:1703.10661. 2017.

[7] Varalakshmi A, Negi A, Krishna S. DataSet generation and feature extraction for Telugu hand-written recognition. International Journal of Computer Science and Telecommunications. 2012;3(3):57-9.

[8] Nasriwala JV. Design and Development of Text Line Segmentation and Recognition of Offline Handwritten Gujarati Text.

[9] Prasad, J.R., Kulkarni, U.V. and Prasad, R.S., 2009, August. Offline handwritten character recognition of Gujrati script using pattern matching. In 2009 3rd international conference on anti-counterfeiting, security, and identification in communication (pp. 611-615). IEEE. https://doi.org/10.1109/icasid.2009.5276999

[10] Z.-H. Zhan, J.-Y. Li, and J. Zhang (2022), ‘‘Evolutionary deep learning: A survey,’’ Neurocomputing, vol. 483, pp. 42–58.

[11] Sadouk, L., Gadi, T. and Essoufi, E.H., 2017, October. Handwritten tifinagh character recognition using deep learning architectures. In Proceedings of the 1st international conference on internet of things and machine learning (pp. 1-11). https://doi.org/10.1145/3109761.3109788

[12] Mohite A, Shelke S. Handwritten Devanagari character recognition using convolutional neural network. In2018 4th International Conference for Convergence in Technology (I2CT) 2018; pp. 1-4. https://doi.org/10.1109/i2ct42659.2018.9057991

[13] Sharma A, Thakkar P, Adhyaru D, Zaveri T. Features fusion based approach for handwritten Gujarati character recognition. Nirma Univ. J. Eng. Technol.(NUJET). 2017;5(2):13-9.

[14] Pareek J, Singhania D, Kumari RR, Purohit S. Gujarati handwritten character recognition from text images. Procedia Computer Science. 2020;171:514-23. https://doi.org/10.1016/j.procs.2020.04.055

[15] Rajyagor B, Rakholia R. Isolated Gujarati handwritten character recognition (HCR) using deep learning (LSTM). In2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT); IEEE. 2021; pp. 1-6. https://doi.org/10.1109/icecct52121.2021.9616652

[16] Suthar SB, Thakkar AR. Dataset Generation for Gujarati Language Using Handwritten Character Images. Wireless Personal Communications. 2024;136(4):2163-84. https://doi.org/10.1007/s11277-024-11369-9

[17] Shukla, D. and Desai, A., 2022. Extraction and recognition of handwritten Gujarati characters and numerals from images using deep learning. In Proceedings of the International e-Conference on Intelligent Systems and Signal Processing: e-ISSP 2020 (pp. 657-669). Springer Singapore. https://doi.org/10.1007/978-981-16-2123-9_51

[18] Limbachiya, K., Sharma, A., Thakkar, P. and Adhyaru, D., 2022. Identification of handwritten Gujarati alphanumeric script by integrating transfer learning and convolutional neural networks. Sādhanā, 47(2), p.102. https://doi.org/10.1007/s12046-022-01864-9

[19] R. Kohavi and F. Provost (1998), ‘‘Glossary of terms,’’ Mach. Learn., vol. 30, pp. 271–274.