IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 31 July 2023, accepted 28 August 2023, date of publication 11 September 2023, date of current version 18 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3314334

RESEARCH ARTICLE

End-to-End Historical Handwritten Ethiopic Text Recognition Using Deep Learning

RUCHIKA MALHOTRA[1] AND MARU TESFAYE ADDIS[2,3]

[1]Department of Software Engineering, Delhi Technological University, Delhi 110042, India
[2]Department of Computer Science and Engineering, Delhi Technological University, Delhi 110042, India
[3]Department of Computer Science, Debre Tabor University, Amhara 251272, Ethiopia

Corresponding author: Maru Tesfaye Addis (marutcomp@gmail.com)

ABSTRACT Recognizing handwritten text is a challenging task, especially for scripts with numerous alphabets and symbols. The Ethiopic script has a vast character set and is used for historical documents in typewritten, handwritten, and hand-printed forms. However, despite its importance as an ancient script, optical character recognition research has not given enough attention to Ethiopic text recognition. In recent years, deep learning (DL) has emerged as a powerful technique for recognizing patterns. In this study, a DL approach is used to recognize historical Ethiopic handwritten texts. The recognition model employs an end-to-end strategy enabling sequential feature extraction and efficient recognition. An attention mechanism coupled with a connectionist temporal classification architecture is the core of this recognition model architecture. In addition, there are seven convolutional neural networks and two recurrent neural networks. We increase the training data using data augmentation techniques to address the data scarcity common in deep learning applications. The experiments include an original training dataset of 79,684 historical handwritten images and an augmented dataset of 10,000 images containing Ethiopic texts. The model used for recognition showed promising results. For "Test Set I" which had 6,150 samples, the character error rate (CER) was 17.95%, and for "Test Set II" which had 15,935 samples, the CER was 29.95%. These outcomes indicate that this approach has the potential to improve the recognition of historical handwritten Ethiopic text.

INDEX TERMS Deep learning, end-to-end learning, ethiopic script, handwritten text recognition, pattern recognition.

I. INTRODUCTION

Ethiopia is incredibly diverse linguistically, with a wide range of languages spoken across its ten states. Each state has its own set of languages, contributing to the country's rich linguistic heritage. The country's linguistic heritage is a testament to its remarkable culture and plays a unifying role among its diverse population. The Ethiopic script is commonly used for writing federal languages such as Amharic and several regional languages including Tigrigna, Awngi, Guragigna, and more. This script plays a significant role in strengthening the cultural and linguistic diversity that thrives in the country.

The Ethiopic script is an old writing system that has been used in Ethiopia and Eritrea. It originated from the South Arabian alphabet [1], [2]. With a history spanning over two thousand years, the Ethiopic script is still in use today, making it one of the few writing systems that has stood the test of time. This script has been used for a variety of purposes, producing historical documents in typewritten, handwritten, and hand-printed formats. For many years, the Ethiopic script has been crucial in safeguarding the linguistic and cultural traditions of Ethiopia and Eritrea. It goes beyond just being a means of communication, as it also serves as a repository of historical records and cultural artifacts. The diversity of languages that use the Ethiopic script adds to its importance, as it showcases the abundant linguistic variety of the region.

In today's world, many government and private organizations are trying to reduce paper usage and move towards digital workflows [3]. However, some institutions such as post offices, banks, and medical institutes often come across handwritten documents in local languages. It's necessary to

The associate editor coordinating the review of this manuscript and approving it for publication was Siddhartha Bhattacharyya.

change these handwritten documents into digital text and customizable formats. There is an increasing demand for efficient optical character recognition (OCR) technology to fulfill this need, which is a crucial component of any text recognition framework [4].

OCR is a system used to transform printed or handwritten text from physical documents into computer understandable formats [3], [5], [6]. Its goal is to transform these documents into electronic versions that can be easily processed and analyzed by computers. Implementing OCR can help organizations improve their document management processes, increase data accessibility, and facilitate efficient information retrieval.

Recognizing handwritten Ethiopic text is challenging due to its vast character set, complex shape, variations in handwriting styles, incomplete strokes, and noise in scanned images. Despite research in pattern recognition for popular scripts, Ethiopic text recognition has not received comparable attention in OCR research [7], [8]. Previous works [9], [10], [11], [12] on Ethiopic script OCR have primarily been based on printed texts, and the recognition of handwritten Ethiopic scripts has remained relatively unexplored [8], [13] due to the scarcity of public research datasets [13]. It is rare to find publicly available historical handwritten datasets, except Belay et al.'s dataset [14], which is typically required for deep learning algorithms [3]. Consequently, recognizing historical handwritten Ethiopic text from image documents poses a unique challenge in OCR. Traditional methods of transcription and analysis can be time-consuming and error-prone, often requiring extensive manual effort. Therefore, there is a need for automated and efficient techniques that can accurately transcribe and interpret the Ethiopic script, facilitating the preservation and understanding of historical documents.

Recent advances in artificial intelligence, specifically the use of deep learning (DL) techniques, have greatly improved pattern recognition. They have proven to be more effective than traditional machine learning (ML) methods [15]. However, these DL techniques require a significant amount of labeled data to work efficiently. Obtaining such data can be difficult and expensive in many cases. To overcome this limitation, image augmentation has emerged as a powerful approach. In computer vision and deep learning, this technique is used to increase the range and size of the training datasets. It enhances the model's ability to generalize and become more resilient. By applying various transformations to the original images, image augmentation replicates real-world scenarios, allowing the model to recognize and handle different image variations more effectively [3].

This advancement in historical handwritten text recognition for the Ethiopic script opens up exciting opportunities for various fields such as linguistic research, historical preservation, and cultural studies. It enables the automated analysis and understanding of historical handwritten texts, which were previously challenging to interpret. Importantly, this study is pioneering in Ethiopic historical handwritten text recognition, being the first of its kind. The study makes significant contributions in the following points:

1) Increasing the dataset: This is done through an augmentation technique applied to the original dataset found in [14]. We have expanded the dataset by adding 10,000 handwritten Ethiopic samples. In order to augment the dataset, we carefully diversified it to ensure that it included a broad range of variations and styles in historical handwritten texts. This enriched dataset has empowered the model with a more extensive and diverse training set, enabling it to learn and recognize historical handwritten texts with enhanced robustness and accuracy.
2) Comprehensive recognition model development: The proposed recognition model combines the strengths of CNN layers for automatic feature extraction, bidirectional LSTM (BLSTM) for sequencing, and CTC loss functions. Through this comprehensive framework, the model is able to automatically extract relevant features, to focus on important sections of the text, and to leverage the temporal value of the text. Furthermore, it facilitates end-to-end training without requiring explicit alignment between the images and labels [16].
3) Experimental evaluation and promising results: Leveraging the augmented dataset, we conducted extensive experiments, including careful hyperparameter selection, to evaluate the performance of the proposed historical handwritten Ethiopic text recognition. Our proposed approach demonstrated potential effectiveness in historical handwritten image text recognition in preliminary results.

The rest of this paper is structured as follows: In Section II, we'll take a look at related research on Ethiopic script recognition. Section III will cover our suggested methodology, including the deep learning architecture and the training process. We'll then move on to the experimental setup and evaluation results in Section IV. Lastly, in Section V, we'll wrap up the paper and highlight potential areas for further research and development in this field.

II. LITERATURE REVIEW

Recognition of handwritten Ethiopic text is an area that has yet to be fully explored in both conventional ML and DL. This is particularly true for historical handwritten text recognition. This section discusses the various methods used to recognize handwritten text. More studies have utilized traditional machine learning approaches, while recent studies have employed deep learning and ensemble approaches.

A. MACHINE LEARNING TECHNIQUES

When studying OCR, machine learning is essential for tasks such as image preprocessing, feature extraction and recognition. Traditional OCR pipelines using machine learning involve various image preprocessing techniques including image denoising, thresholding, and morphological operations. These techniques improve image quality and eliminate noise or artifacts in preparation for further analysis [17], [18].

After preprocessing the images, various feature extraction algorithms are utilized to capture distinctive information from them. These algorithms may involve techniques such as Histogram of Oriented Gradients (HOG), Scale-Invariant Feature Transform (SIFT), or Local Binary Patterns (LBP) [19]. The purpose of these algorithms is to extract relevant features that encode the texture, shape, or structural characteristics of the text in the image [20], [21].

Once the features have been extracted, recognition is carried out using a range of well-known machine learning algorithms such as random forests, support vector machines, k-nearest neighbors, or multilayer perceptrons (MLPs). These algorithms use the extracted features to train models that can accurately recognize text in images.

For optimal OCR results, choose preprocessing, feature extraction, and ML algorithms that fit the task and dataset characteristics. By using ML techniques, researchers can automate text extraction from images, which enables a range of applications, including document digitization, text translation, and information retrieval. These advancements have opened new doors for linguistic research, historical preservation, and cultural studies, as they allow for automated analysis and comprehension of historical handwritten texts in various scripts, such as the Ethiopic script.

B. END-TO-END LEARNING

End-to-End (E2E) learning is an approach to machine learning that aims to simplify the process of solving a task by combining all the necessary stages into a single model [22]. E2E learning eliminates the need for handcrafted features or intermediate representations. Instead, the model can learn directly from raw input data to produce the desired output. Recently, there has been a surge in the use of deep learning and neural networks [23]. These models, including CNN and RNN, have shown great success [24] in fields such as computer vision, natural language processing, speech recognition, and pattern recognition [25]. One major benefit of E2E learning is its simplicity in design. It eliminates the need for complicated feature engineering, making the overall system design more straightforward. Moreover, E2E learning models are adaptable and can conform to various domains and tasks. They can learn pertinent features and representations from raw data, capturing complex patterns and dependencies effectively [26].

It's essential to recognize that E2E learning models have specific limitations. These models require a significant amount of labeled training data to perform effectively in different situations, which can be an expensive and time-consuming process. Moreover, the integrated nature of E2E models can make it difficult to repurpose or adjust specific components for various tasks [27]. E2E learning is a powerful technique that helps models generate desired outputs from raw data. It requires labeled data and may be difficult to interpret or modularize.

C. HANDWRITTEN ETHIOPIC TEXT RECOGNITION

Recognizing handwritten Ethiopic text is challenging due to the intricate nature of the Ethiopic script and the limited availability of annotated datasets. Despite this, some studies in recent years have suggested both conventional ML and DL techniques to recognize handwritten Ethiopic text. Here are some examples:

Alemu and Fuchs [2] published a research paper outlining their approach to recognizing handwritten Amharic bank checks. This study is significant as it is the first of its kind to tackle this challenge. The main obstacle in recognizing handwritten Amharic text is that different writers may use different writing styles for the same numerals. HMRF can be used to extract relevant features from handwritten characters in order to address this issue. The HMRF algorithm can model context-dependent entities based on the concept of Markov Random Field (MRF) theory, which considers both local and global interactions. HMRF is not typically used in handwritten recognition because of its extensive computational demands. However, the authors considered the possibility of acceptable time and space consumption in their approach, making a valuable contribution to handwriting recognition, particularly in the Amharic language. Their recognition model was evaluated using training images consisting of 7,240 characters and a testing image set consisting of 713 characters. They incorporated contextual information in their approach, leading to a significant increase in accuracy from 89.06% to an impressive 99.44%. It is worth noting that their study did not consider the influence of prior probability, which could have an impact on the obtained results.

A study conducted by Assabie and Begun [28] focused on recognizing offline handwritten Amharic words. The goal was to tackle the challenges posed by the large number of character classes and the complexity of Amharic script. To achieve this, they used a hidden Markov model (HMM) for recognizing unconstrained handwritten Amharic words. The authors' recognition pipeline involved feature extraction and recognition steps. They used direction field image computation and segmentation techniques at the text line, word, and pseudo-character levels to capture the unique structural characteristics of the Amharic script. According to their experiment, the recognition rate was 76% for good-quality image categories and 53% for poor-quality images. However, the authors acknowledged that further improvements could be achieved by enhancing the extraction of structural features and incorporating language models into the HMM framework. Future work could focus on refining the extraction of structural features, such as curves, loops, and junctions, and integrating language models, such as n-grams or recurrent neural networks, into the HMM-based recognition system. By addressing these aspects, the recognition accuracy of Amharic handwritten words can be further improved, which contributes to the development of more robust and accurate systems for Amharic language processing.

In their research, Tamir [29] explored handwriting recognition of Amharic characters using CNN. The goal of this study was to overcome the challenges faced in managing ancient handwritten documents in Amharic, which are prone to deterioration over time. An automatic approach to handwritten text recognition was proposed, utilizing a CNN design consisting of two convolutional layers to classify handwritten Amharic characters. Batch Normalization and Activation layers were sequentially implemented after each convolutional layer. In addition, Max-pooling was executed after the activation layer of the second convolutional layer. Finally, the output was flattened to be fully connected to the final layer, which was responsible for character classification. To conduct the research, data was collected from about 130 individuals, resulting in a dataset of 30,446 characters. Of this dataset, 27,413 characters were used for model training, while the remaining 3,033 were reserved for testing and evaluating the model's performance. Although the author provided information on the training accuracy and loss, it is important to also have validation accuracy and loss metrics to determine the model's ability to generalize. These metrics allow for assessing how well the model performs on new data and provide insights into its overall effectiveness. It is recommended that the author includes both validation accuracy and validation loss in their report for better research. This will help to evaluate the model's performance comprehensively and determine its capacity to identify new and unseen handwritten Amharic characters. Furthermore, it would be beneficial to analyze and interpret the results to highlight the strengths and limitations of the proposed method and identify areas for future research and improvement.

In a study by Agegnehu et al. [30], deep learning was employed to identify Amharic punctuation marks and handwritten digits using a CNN architecture. The dataset comprised of 5,800 images sourced from 100 handwriting samples, with a testing accuracy of 70.04% achieved. Additionally, research has also investigated Ge'ez digits in contexts other than printed and handwritten forms.

Most researchers have primarily concentrated on recognizing English numerals, and the lack of openly accessible Ge'ez digit datasets has hindered extensive research in this field. Nonetheless, Nur et al. [13] have made progress in Ge'ez digit recognition by creating a recognition model with better accuracy and an experimental dataset consisting of 51,952-digit images written by 524 people. Their proposed CNN model has achieved a recognition accuracy of 96.21%. The authors have suggested that future research should explore various deep-learning approaches for multi-digit recognition.

Our research is focused on recognizing historical handwritten Ethiopic text in images by utilizing E2E approach. To achieve this, we acknowledge the importance of deep convolutional recurrent neural networks (CRNNs) and connectionist temporal classification (CTC) for the efficient application of end-to-end learning. CNNs have proven crucial in image processing and pattern recognition, especially in tasks involving automatic feature extraction from images [10].

III. METHODS

This section presents the E2E historical handwritten Ethiopic image text recognition using DL. This study aims to develop a recognition model that accurately recognizes handwritten text from an image. An explanation of the model's architecture and the dataset used for training and evaluating the model is provided.

A. THE DATASET PREPARATION

During this research, we realized that having a large dataset is crucial for developing and testing deep learning models. To overcome the challenge of limited data, we needed a dataset that was specially designed for our deep learning problem.

The dataset used in this study called "HHD-Ethiopic," which was created by Belay et al. [14]. This dataset contained 79,684 images of handwritten historical documents that featured 306 unique Ethiopic characters. We divided the data into training, validation, and testing subsets. The training set had 57,374 samples, while the remaining samples were kept for testing.

To ensure accurate assessment, we created a validation dataset using 10% of the training dataset. The testing dataset was divided into two parts: "Test Set I" and "Test Set II." Test Set I had 6,375 images randomly chosen from the training set while Test Set II contained 15,395 images from 18th-century manuscripts. The original dataset is available on the GitHub repository (https://github.com/bdu-birhanu/HHD-Ethiopic/tree/main/Dataset). We faced the challenge of having a small dataset for deep learning tasks. To increase its size, we used data augmentation techniques. Despite resource limitations, we were able to add around 10,000 augmented images. Sample augmented image is shown in **FIGURE 1**. By using image augmentation, deep learning models can learn effectively from limited labeled data. This leads to better performance and more reliable predictions. This technique has become an essential tool in overcoming data scarcity and maximizing the potential of DL algorithms. These algorithms excel in various applications across different domains.

Image augmentation techniques involve applying operations like rotations, shifting, zooming, shearing, flipping, and noise injection. These transformations create new training examples that are variations of the original images while preserving their semantic content. By introducing such variations, the model becomes more robust to changes in factors that can be encountered during inference on real-world data.

During the training process, augmented images are commonly utilized to offer a wider range of examples for the model to learn from. This technique helps to prevent overfitting and enhances the model's capacity to perform well

FIGURE 1. Sample images from the HHD-Ethiopic database (a) represents the original image and (b) represents the augmented image.

on unseen data. Image augmentation is particularly useful when the available training dataset is limited, as it enables the creation of additional training samples without the need for manual data collection or annotation.

The primary objective of expanding the dataset through data augmentation was to enhance the model's performance and generalization abilities. By providing a larger and more diverse set of examples for the model to learn from, we aimed to improve its ability to recognize and handle various patterns and variations in the data. To illustrate this, sample augmented images can be seen in ***FIGURE 1***, which depicts how the dataset was enriched to facilitate better learning and adaptation in our deep learning approach.

B. ARCHITECTURE OF THE PROPOSED MODEL

This study utilizes the advantages of several deep learning algorithms by combining them, namely CNN, BLSTM, Attention mechanisms, and CTC. Each algorithm has a crucial role in different recognition stages. By bringing these algorithms together, we create a comprehensive and effective approach to recognizing handwritten text. Our careful selection and use of these algorithms contribute to the success of this study and enable more accurate and reliable recognition. Below is a detailed explanation of each algorithm.

1) CNN LAYERS

Our study aimed to extract deeper features from handwritten text images using CNNs, which are known for their exceptional ability to identify patterns and significant information within images [31]. Previous research has successfully applied CNNs to various tasks [32], and we utilized them to accurately recognize the distinctive characteristics of handwritten text. The CNN's parameters were carefully selected and fine-tuned to ensure the extracted features were precise.

To represent the mathematical effectiveness of CNNs in feature extraction, let's consider the input handwritten text image I with pixel values denoted by $I(x, y)$, where x and y are the spatial coordinates of the image. A convolutional layer applies a set of learnable filters (also known as kernels) to the input image. The output of the convolutional layer can be computed in "(1)".

$$O(i,j) = \sum\nolimits_{m=1}^{M} \sum\nolimits_{n=1}^{N} I(i+m, j+n) \cdot W(m,n) \qquad (1)$$

where:

- $O(i, j)$ represents the output feature map at the spatial location (i, j), $W(m, n)$ denotes the learnable weights of the filter at the position (m, n), M and N are the dimensions of the filter.

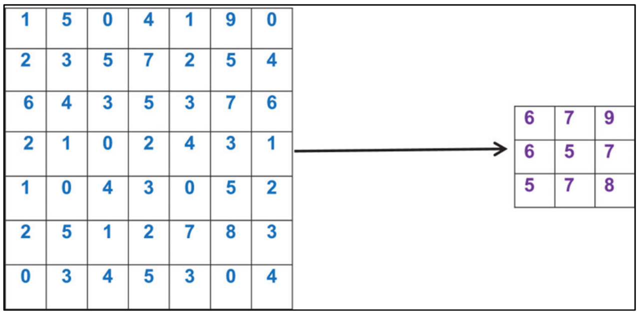
After the convolution operation, an activation function such as Rectified Linear Unit (ReLU) is applied element-wise to introduce non-linearity using "(2)".

$$F(i,j) = ReLU(O(i,j)) \qquad (2)$$

The process of pooling (commonly max-pooling) reduces the spatial dimensions of the feature maps, which helps in reducing computational complexity and controlling overfitting. The pooled output is calculated in "(3)".

$$P(i,j) = max(F(s \cdot i, s \cdot j)) \qquad (3)$$

where $P(i, j)$ represents the pooled output, and s is the stride of the pooling operation. The max pooling results with a 3 × 3 filter with strides 2 are shown in **FIGURE 2**.



1	5	0	4	1	9	0
2	3	5	7	2	5	4
6	4	3	5	3	7	6
2	1	0	2	4	3	1
1	0	4	3	0	5	2
2	5	1	2	7	8	3
0	3	4	5	3	0	4

6	7	9
6	5	7
5	7	8

FIGURE 2. Max pool with 3 × 3 filter and 2 strides [31].

The mathematical representations and fine-tuning of CNN parameters play a crucial role in this feature extraction process, making our study a significant contribution to the field of image recognition and analysis.

2) BLSTM LAYERS

In this study, we leverage the power of BLSTM networks to model the difficult sequential dependencies and temporal dynamics inherent in handwritten text. The inclusion of BLSTM layers empowers the network to efficiently capture contextual information [32], thereby significantly enhancing the accuracy of recognition [4].

"Equation (4)-(7)" is used to compute the mathematical formulation of a BLSTM unit. At each time step t, the BLSTM unit takes as input the hidden state h_{t-1} from the previous time step, the current input x_t (which can be a vector representation of a character or a pixel in the

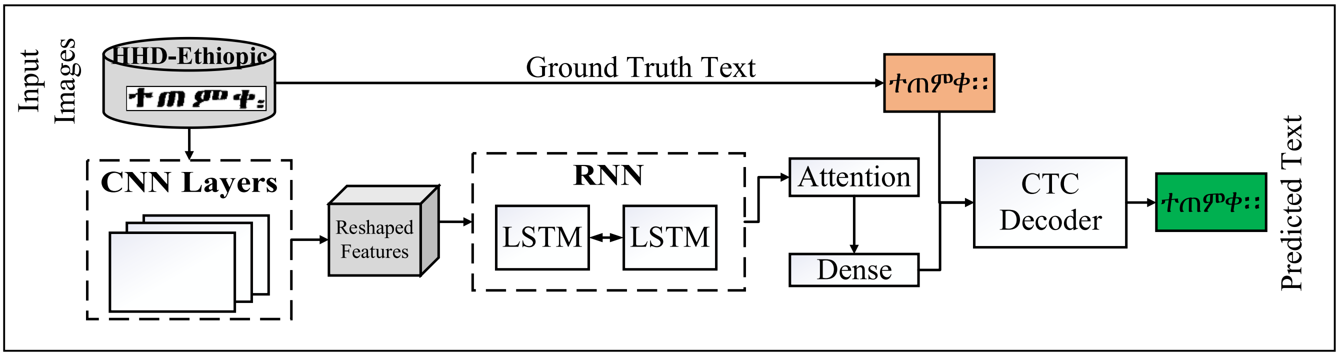


FIGURE 3. The general architecture of the proposed historical handwritten Ethiopic text recognition.

handwritten image), and computes the following intermediate values.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (4)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (5)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (6)$$

$$g_t = tanh(W_g \cdot [h_{t-1}, x_t] + b_g) \quad (7)$$

where:

- σ represents the sigmoid activation function, tanh denotes the hyperbolic tangent activation function, W_f, W_i, W_o, and W_g are the learnable weight matrices, b_f, b_i, b_o, b_g are the learnable bias vectors, $[h_{t-1}, x_t]$ denotes the concatenation of h_{t-1} and x_t along the feature dimension. The intermediate values f_t, i_t, o_t, and g_t are used to update the cell state c_t and the hidden state h_t of the BLSTM unit as in "(8)" and "(9)".

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (8)$$

$$h_t = o_t \odot tanh(c_t) \quad (9)$$

where $\odot$ denotes the element-wise multiplication.

The output of the BLSTM layer is then used for further processing, such as classification in the case of handwritten text recognition.

By incorporating BLSTM layers into our model, we equip it with the ability to comprehend the underlying structure and context of the handwritten text comprehensively. This deep understanding of sequential information facilitates improved accuracy in recognition tasks, making our study a valuable contribution to the field of handwritten text analysis and understanding.

3) ATTENTION LAYERS

Attention mechanisms have proven to be effective in enhancing the model's ability to focus on relevant regions within the handwritten text. By dynamically weighting different parts of the sequence during recognition, the model can allocate its attention to the most salient features and characters, thereby improving recognition accuracy, especially when dealing with inter-class similarity and structural complexity challenges.

The attention mechanism can be mathematically described as follows:

Let $H = \{h_1, h_2, \ldots, h_T\}$ be the set of hidden states produced by the encoder, where T is the length of the input sequence. These hidden states capture valuable information about the handwritten text at each time step. The attention mechanism generates a set of context vectors $C = \{c_1, c_2, \ldots, c_T\}$ that represent the weighted combinations of the encoder's hidden states. The context vector c_t at the time step t is computed as a weighted sum of all hidden states h_i with attention weights $a_{t,i}$ in "(10),".

$$c_t = \sum_{i=1}^{T} a_{t,i} \cdot h_i \quad (10)$$

The attention weights $a_{t,i}$ are typically calculated using a scoring function that measures the relevance of the input at the time step i to the output at the time step t. One commonly used scoring function is the dot product as represented in "(11)".

$$e_{t,i} = h_t^T \cdot h_i \quad (11)$$

where $e_{t,i}$ represents the score between the hidden state at the time step t and the hidden state at the time step i. The scores $e_{t,i}$ are then normalized using the SoftMax function to obtain the attention weights calculated in "(12)".

$$a_{t,i} = \frac{exp(e_{t,i})}{\sum_{j=1}^{T} exp(e_{t,j})} \quad (12)$$

By calculating the attention weights for each time step, the model can effectively focus on the most relevant regions within the handwritten text during recognition. This adaptability allows the model to emphasize distinctive features and characters, thereby overcoming challenges posed by inter-class similarity and structural complexity.

The context vectors C are then combined with the decoder's hidden states to generate the final output sequence during the decoding (inference) process.

In summary, the integration of attention mechanisms enables the model to dynamically allocate its focus on salient regions within the handwritten text, leading to improved recognition accuracy, particularly in scenarios where class

TABLE 1. The proposed historical handwritten Ethiopic text recognition layers and their Hyper-parameter values.

S. No	Layers	Configuration
1.	Input	Rows:48, Columns: 368, Channels:1
2.	Conv1	Feature:32, kernel:(3,3), activation: ReLU, padding: same
3.	MaxPooling1	Pooling size: (2,1), strides:2
4.	Conv2	Feature:32, kernel:(3,3), activation: ReLU, padding: same
5.	MaxPooling2	Pooling size: (2,2)
6.	Conv3	Feature:32, kernel:(3,3), activation: ReLU, padding: same
7.	Conv4	Feature:32, kernel:(3,3), activation: ReLU, padding: same
8.	MaxPooling3	Pooling size: (2,2)
9.	Normalize	Batch Normalization
10.	Conv5	Feature:32, kernel:(3,3), activation: ReLU, padding: same
11.	Conv6	Feature:32, kernel:(3,3), activation: ReLU, padding: same
12.	Normalize	Batch Normalization
13.	Conv7	Feature:128, kernel:(2,2), activation: ReLU
14.	BLSTM1	RNN Units: 128, droout:0.25, return sequence: True
15.	BLSTM2	RNN Units: 128, droout:0.25, return sequence: True
16.	Attention	Feature: 91, activation: tanh, SoftMax
17.	Dense	Units: no_class+1, activation: SoftMax
18.	Output + CTC	max_len:46, input length: Label length:

similarities and structural intricacies are prominent. This mathematical formulation of attention mechanisms highlights their significant impact on the performance of the model in handwritten text recognition tasks.

4) CTC LAYER

CTC is a valuable technique used in sequence recognition tasks, particularly for applications like handwritten text recognition. CTC enables the model to learn from sequences of variable length without requiring explicit alignment between input images and their corresponding labels. This makes it a powerful tool for end-to-end training and decoding. The CTC loss function is employed to train the model. Given an input sequence of feature vectors (representing the handwritten text image) denoted by $X = \{x_1, x_2, \ldots x_T\}$, and the corresponding label sequence denoted by $Y = \{y_1, y_2, \ldots y_T\}$ where T and U are the lengths of the input sequence and label sequence, respectively, the CTC loss is computed using "(13)".

Let's define S as the set of all possible labels, including a special "blank" symbol denoted by Ø. The CTC loss function $L(X, Y)$ is the negative log-likelihood of the correct label sequence given the input sequence X.

$$L(X, Y) = -logp(Y \mid X) \tag{13}$$

The probability $p(X|Y)$ is computed by summing over all valid alignments between the input sequence X and the label sequence Y in "(14)".

$$p(Y \mid X) = \sum\nolimits_{align \in A(X,Y)} \prod_{t=1}^{T} p(y_{align(t)} \mid x_t) \tag{14}$$

where $A\,(X, Y)$ is the set of valid alignments, considering the blank symbol and allowed repetitions and removals of labels during the alignment process.

During training, the CTC loss is minimized using gradient-based optimization techniques like stochastic gradient descent (SGD) or Adam to update the model's parameters, allowing it to learn to recognize handwritten text effectively.

Furthermore, during decoding (inference), CTC enables the model to produce the most probable label sequence directly from the input sequence without requiring explicit alignment. This is achieved through a decoding algorithm, such as the beam search algorithm, which explores the possible label sequences and selects the most likely output sequence.

The general architecture for historical handwritten Ethiopic text recognition is illustrated in **FIGURE 3**. The diagram depicts the flow of the recognition model, starting with the input image obtained from the "HHD-Ethiopic" image database. The input image dimensions, in our case, 48 by 368 pixels, are then passed through a series of CNN layers. The model architecture consists of seven convolutional layers along with ReLU activation function for each, three max-pooling, and two batch normalization, as depicted in **TABLE 1**. The extracted features from the CNN layers are then passed to the next two BLSTM layers for sequence modeling. Each LSTM layer is constructed with 128 hidden units. Next, the attention mechanism is applied by concatenating the LSTM layers. A fully connected or dense layer with Soft Max activation processes features and prepares them for classification. The number of units in the dense layer is determined based on the number of classes or categories in the recognition task. A total of 306 alphabet symbols are included in this study. Therefore, there are 307 classes, including the blank CTC space. The model is compiled with the Adam optimizer and CTC loss. Finally, the CTC decoder generates the predicted result by cross-checking the ground truth labels with the alphabet or character set. The CTC decoder ensures that the recognition model can handle sequences of variable lengths and directly learn from input-output alignment without explicit alignment.

C. PERFORMANCE EVALUATION METRICS

Our proposed model's performance is assessed using character error rate (CER). The CER is calculated using equation (1).

$$CER = \left(\frac{(I + D + S)}{GT}\right) * 100, \tag{15}$$

Here, I, D, and S represent the number of character insertions, deletions, and substitutions respectively. GT denotes the total number of characters in the ground truth text.

The CER provide valuable insights into the error rates of the recognition model. By evaluating these metrics, we can measure the effectiveness of our proposed model in accurately transcribing characters, and identify areas where improvements may be required.

We used additional evaluation metrics, alongside our primary CER metric, to provide a comprehensive analysis of the model's performance. These metrics, including precision, recall, and F1-score, focus on evaluating the model's classification accuracy, specifically in identifying characters and their positions in the sequence.

To better understand these metrics, let's explain the parameters used in generating the classification report. TP (true positive) refers to the number of positive instances correctly predicted, while FP (false positive) refers to the number of positive instances predicted incorrectly. Similarly, TN (true negative) represents the number of negative instances correctly predicted, and FN (false negative) signifies the number of negative instances predicted incorrectly. "Equations (16)", "(17)", and "(18)" are used to calculate precision, recall, and F1-score respectively.

$$P = \frac{TP}{(TP + FP)}, \tag{16}$$

$$R = \frac{TP}{(TP + FN)}, \tag{17}$$

$$F = \frac{2PR}{R + P}, \tag{18}$$

Here, the precision value is represented by P, recall value is represented by R, and F-score is represented by F.

IV. EXPERIMENTAL SETUP, RESULTS, AND DISCUSSION

We conducted extensive experiments utilizing deep learning techniques to demonstrate the efficacy of the proposed historical handwritten text recognition model. In this section, we present experimental setup and a detailed overview of the performance evaluation measures employed, followed by a description of the experimental setup. We then present the results and engage in a thorough discussion of the experiments conducted using two testing datasets. Furthermore, we discuss the techniques employed for selecting optimal parameters for the designed architectural model and provide the results obtained from these optimal parameter settings. Finally, we perform an in-depth error analysis of the results.

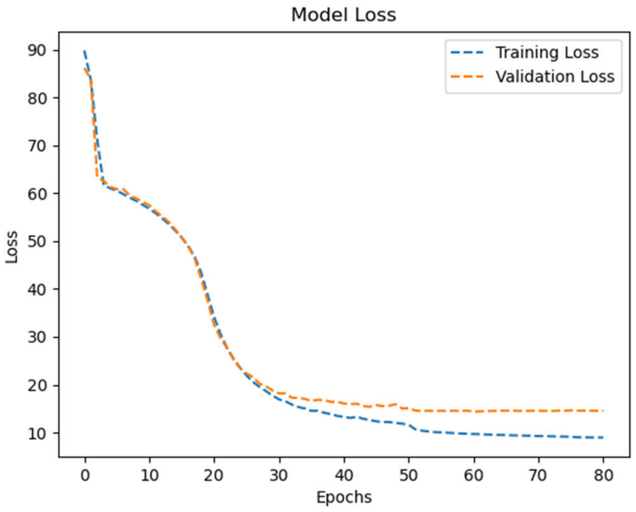


FIGURE 4. Training and validation loss vs epoch graph for the first model.

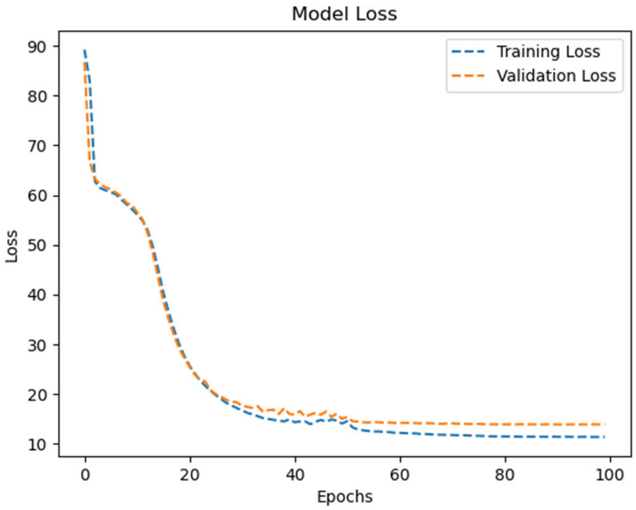


FIGURE 5. Training and validation loss vs epoch graph for the second model.

A. EXPERIMENTAL SETUP

The proposed model was implemented in Python using the Keras framework with TensorFlow as the backend. This combination allowed for the efficient development and training of the model. Powerful GPUs were utilized to expedite the training process and take advantage of accelerated computation. This was made possible through Kaggle notebooks, eliminating expensive hardware. Kaggle, a cloud-based

TABLE 2. **CER test results of the first model.**

Test Set Type	Samples	CER
Test Set I	6,150	18.33%
Test Set II	15,935	30.69%

TABLE 3. **CER test results of the second model.**

Test Set Type	Samples	CER
Test Set I	6,150	17.95%
Test Set II	15,935	29.95%

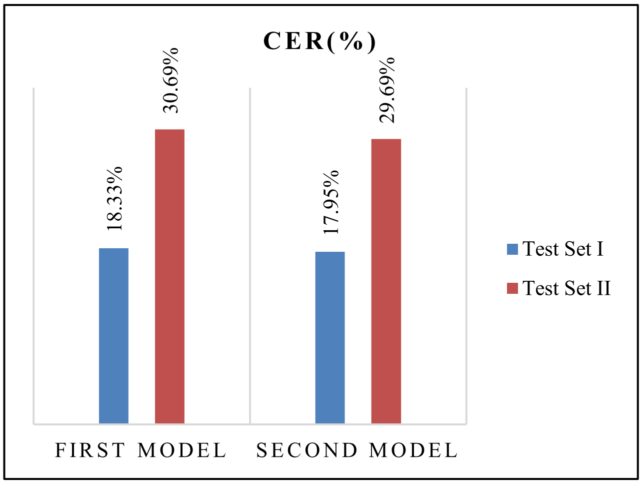


FIGURE 6. **CER comparison of the first and the second models.**

TABLE 4. **Precision, recall, F-score results of the better model.**

Test set type	Precision	Recall	F-score	Support
Test set I	0.9014	0.8780	0.8895	89780
Test set II	0.8488	0.8011	0.8243	280917

service, offers remote code execution from any location with an internet connection. Its user-friendly interface and comprehensive support for popular data science libraries make it an excellent platform for deep learning projects. We streamlined the development process by utilizing Kaggle's capabilities and efficiently training the proposed model.

B. RESULTS

In this research, we created two recognition models with 100 epochs and a batch size of 64, each selected numerically. The first model utilized the original historical handwritten images, while the second model incorporated augmentation by adding 10,000 images generated from the original dataset. In the following results section, we present the outcomes achieved by both historical handwritten text recognition models for the Ethiopic script. **FIGURE 4** and **FIGURE 5** showcase the training and validation losses plotted against the number of epochs for the first and second models, respectively. These graphs provide valuable insights into each model's learning progress and performance throughout the training process.

TABLE 2 and **TABLE 3** displays the CER values of the first and the second model respectively. The number of samples in each testing set also described.

1) ERROR RESULTS

In **FIGURE 7**, we present the error results from the 18th century testing dataset for the second recognition model. The text on the left shows the actual ground-truth texts, while the text on the right displays the predicted texts generated by our recognition model. Characters that were incorrectly recognized by the recognition model are highlighted in green, while characters that were present in the ground-truth texts but not identified as characters in the predictions are shown in blue. In the predicted text, characters that are not recognized correctly are colored in red.

2) CORRECT RESULTS

FIGURE 8, displays sample images which are recognized by correctly by the proposed models.

C. DISCUSSION

This study compares two models for recognizing handwritten Ethiopic text. The first model uses the "HDD-Ethiopic" dataset, while the second model uses augmented images. The results show that the model trained with augmented images performs better in terms of CER on both testing datasets, "test set I" and "test set II", as shown in **TABLE 2** and **TABLE 3**. The first model achieved a CER of 18.33% and 30.69% for "test set I" and "test set II", respectively. The second model achieved a CER of 17.95% and 29.95% for "test set I" and "test set II", respectively. The reason for the second model's better performance is the use of augmented images during training. This expands the training data, improves model generalization, and enhances the ability to recognize diverse variations in handwritten Ethiopic text. As a result, the model trained with augmented images achieves superior results.

We analyzed the superior model or second model and presented its precision, recall, and f-score results in **TABLE 4** for both testing datasets. The second model performed better in all metrics for test set I, confirming its superiority. However, test set II had more instances of highly confused characters, despite having a larger support.

The character "፡" is frequently used in this dataset. Because this character is used in every word to separate. So, it has better recognition results for precision (0.9973), recall (0.9955), and F-score (0.9964).

Ethiopic characters often share similar or approximate shapes, which can lead to confusion during recognition. For example, as we observed in **FIGURE 7**, the character "ጎ" is recognized as "ን". When we see these two characters almost they have similar shape on the top part. The character "ወ" is recognized as "ጠ". Here from the ground truth text "ወ" is connected but the predicted character "ጠ" doesn't

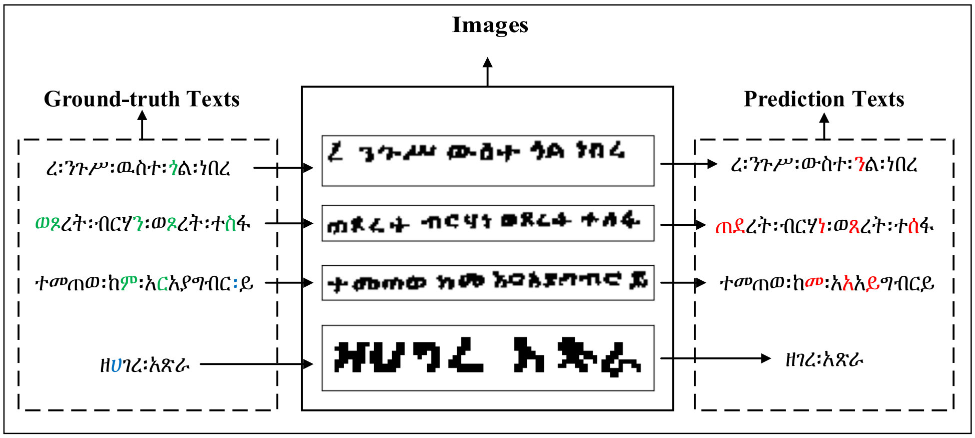


FIGURE 7. Sample error results from the second model using "Test Set II".

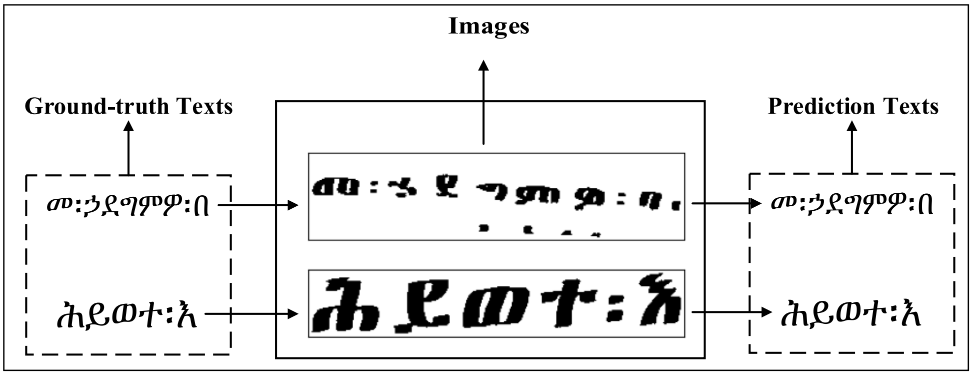


FIGURE 8. Sample correctly recognized results in both models.

have connection at the bottom of the character. This is the minor difference in terms of shape. Other characters also misrecognized as "ጰ" to "ደ", "ን" to "ነ" and "ሰ" to "ሰ".

Some texts are fully recognized correctly with in the sequence. We can observe from **FIGURE 8**. Here all the characters inside the sequence are fully recognized.

V. CONCLUSION

In summary, the recognition of Ethiopic OCR is an important area of research with numerous applications. Therefore, a recognition model is necessary for this purpose. This study presents an end-to-end learning model that is specifically designed to recognize historical handwritten Ethiopic text. With the use of deep learning techniques, our model automatically extracts relevant features from input images and effectively captures the sequential nature of the text. Additionally, the integration of attention mechanisms and a CTC-based loss function enables the model to concentrate on crucial regions of the input, allowing for end-to-end training without requiring explicit alignment between images and labels. Our proposed approach was found to be effective, as demonstrated by experimental results.

To improve our recognition system, we can expand the dataset, explore alternative deep learning architectures, and test it in other languages. Using a more diverse dataset will enhance performance and reduce errors. With ongoing research, we can achieve more accurate recognition, benefiting various applications.

ACKNOWLEDGMENT

The authors would like to thank the faculty members, staff, and students of Delhi Technological University for their invaluable guidance and support throughout the completion of this manuscript. Their continuous assistance and expertise have played a crucial role in the success of this research project and also would like to thank unwavering commitment to fostering academic excellence and their contributions to their growth as a researchers. Additionally, "ChatGPT (GPT-3.5)" and "Word tune" have been extremely helpful in the development of this study.

REFERENCES

[1] R. Meyer, "The Ethiopic script: Linguistic features and socio-cultural connotations," *Oslo Stud. Lang.*, vol. 8, no. 1, pp. 137–172, Feb. 2017.

[2] W. Alemu and S. Fuchs, "Handwritten Amharic bank check recognition using hidden Markov random field," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop*, Jun. 2003, p. 28.

[3] P. Goel and A. Ganatra, ''Handwritten Gujarati numerals classification based on deep convolution neural networks using transfer learning scenarios,'' *IEEE Access*, vol. 11, pp. 20202–20215, 2023.

[4] Y. S. Chernyshova, A. V. Sheshkus, and V. V. Arlazarov, ''Two-step CNN framework for text line recognition in camera-captured images,'' *IEEE Access*, vol. 8, pp. 32587–32600, 2020.

[5] J. Memon, M. Sami, R. A. Khan, and M. Uddin, ''Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR),'' *IEEE Access*, vol. 8, pp. 142642–142668, 2020.

[6] T. Nasir, M. K. Malik, and K. Shahzad, ''MMU-OCR-21: Towards end-to-end Urdu text recognition using deep learning,'' *IEEE Access*, vol. 9, pp. 124945–124962, 2021.

[7] E. Y. Obsie, H. Qu, and Q. Huang, ''Amharic character recognition based on features extracted by CNN and auto-encoder models,'' in *Proc. 13th Int. Conf. Comput. Model. Simul.*, Melbourne VIC, Australia, Jun. 2021, pp. 58–66.

[8] B. Belay, T. Habtegebrial, M. Liwicki, G. Belay, and D. Stricker, ''Factored convolutional neural network for Amharic character image recognition,'' in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 2906–2910.

[9] M. Meshesha and C. V. Jawahar, ''Recognition of printed Amharic documents,'' in *Proc. 8th Int. Conf. Document Anal. Recognit. (ICDAR)*, 2005, pp. 784–788. [Online]. Available: https://ieeexplore.ieee.org/ielx5/10526/33307/01575652.pdf?tp=&arnumber=1575652&isnumber=33307&ref=

[10] B. H. Belay, T. A. Habtegebrial, and D. Stricker, ''Amharic character image recognition,'' in *Proc. IEEE 18th Int. Conf. Commun. Technol. (ICCT)*, Chongqing, China, Oct. 2018, pp. 1179–1182.

[11] R. Malhotra and M. T. Addis, ''Ethiopic base characters image recognition using LSTM,'' in *Proc. 2nd Int. Conf. Comput. Methods Sci. Technol. (ICCMST)*, Mohali, India, Dec. 2021, pp. 94–98.

[12] D. Addis, C.-M. Liu, and V.-D. Ta, ''Printed ethiopic script recognition by using LSTM networks,'' in *Proc. Int. Conf. Syst. Sci. Eng. (ICSSE)*, Jun. 2018, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/ielx7/8500054/8519965/08519972.pdf?tp=&arnumber=8519972&isnumber=8519965&ref=

[13] M. Ali Nur, M. Abebe, and R. S. Rajendran, ''Handwritten geez digit recognition using deep learning,'' *Appl. Comput. Intell. Soft Comput.*, vol. 2022, pp. 1–12, Nov. 2022.

[14] B. H. Belay, ''HHD-Ethiopic: A historical handwritten dataset for Ethiopic OCR with baseline models and human-level performance (revision 50C1E04),'' Tech. Rep., 2023.

[15] S. Aly and A. Mohamed, ''Unknown-length handwritten numeral string recognition using cascade of PCA-SVMNet classifiers,'' *IEEE Access*, vol. 7, pp. 52024–52034, 2019.

[16] Y. Zhu, Z. Xie, L. Jin, X. Chen, Y. Huang, and M. Zhang, ''SCUT-EPT: New dataset and benchmark for offline Chinese text recognition in examination paper,'' *IEEE Access*, vol. 7, pp. 370–382, 2019.

[17] M. Sonka, V. Hlavac, and R. Boyle, *Image Processing, Analysis, and Machine Vision*, T. L. Anderson, Ed. 2014, p. 930.

[18] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, *Digital Image Processing Using MATLAB*. Upper Saddle River, NJ, USA: Prentice-Hall, 2004, p. 302.

[19] D. B. Honnaraju, M. Meghana, D. S. Sanjana, N. S. Nisarga, and H. R. Nikhil, ''Sign language recognition using deep learning (CNN) and SVM,'' *Int. Res. J. Modernization Eng. Technol. Sci.*, vol. 5, no. 5, pp. 6479–6483, 2023.

[20] N. Dalal and B. Triggs, ''Histograms of oriented gradients for human detection,'' in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Diego, CA, USA, Mar. 2005, pp. 886–893.

[21] D. G. Lowe, ''Distinctive image features from scale-invariant keypoints,'' *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[22] Y. LeCun, Y. Bengio, and G. Hinton, ''Deep learning,'' *Nature*, vol. 521, pp. 436–444, May 2015.

[23] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[24] M. G. Gurmu, ''Offline handwritten text recognition of historical Ge'ez manuscripts using deep learning techniques,'' M.S. thesis, Inf. Sci., Jimma Univ., Jimma, Ethiopia, 2021.

[25] N. Mungoli, ''Adaptive feature fusion: Enhancing generalization in deep learning models,'' *Int. J. Comput. Sci. Mobile Appl.*, vol. 11, no. 3, pp. 1–11, 2023.

[26] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, ''Gradient-based learning applied to document recognition,'' *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[27] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, ''OverFeat: Integrated recognition, localization and detection using convolutional networks,'' in *Proc. 2nd Int. Conf. Learn. Represent.*, Apr. 2014, pp. 14–16.

[28] Y. Assabie and J. Bigun, ''HMM-based handwritten amharic word recognition with feature concatenation,'' in *Proc. 10th Int. Conf. Document Anal. Recognit.*, 2009, pp. 961–965.

[29] K. Tamir, ''Handwritten Amharic characters recognition using CNN,'' in *Proc. IEEE AFRICON*, Accra, Ghana, Sep. 2019, pp. 1–4.

[30] M. Agegnehu, G. Tigistu, and M. Samuel, ''Offline handwritten Amharic digit and punctuation mark script recognition using deep learning,'' in *Proc. 2nd Deep Learn. Indaba-X Ethiopia Conf.*, Adama, Ethiopia, Jan. 2022.

[31] H. T. Weldegebriel, H. Liu, A. U. Haq, E. Bugingo, and D. Zhang, ''A new hybrid convolutional neural network and extreme gradient boosting classifier for recognizing handwritten Ethiopian characters,'' *IEEE Access*, vol. 8, pp. 17804–17818, 2020.

[32] B. Shi, X. Bai, and C. Yao, ''An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2298–2304, Nov. 2017.

RUCHIKA MALHOTRA received the master's and Ph.D. degrees in software engineering from the University School of Information Technology, Guru Gobind Singh Indraprastha University, Delhi, India. She is currently a Professor with the Department of Software Engineering, Delhi Technological University, Delhi. She has published more than 200 research papers in international journals and conferences. Her research interests include software testing, improving software quality, statistical and adaptive prediction models, software metrics, neural nets modeling, and the definition and validation of software metrics.

MARU TESFAYE ADDIS received the bachelor's degree in computer science and information technology from Arba Minch University, Ethiopia, in 2010, and the master's degree in computer science from Bahir Dar University, Ethiopia, in 2017. He is currently pursuing the Ph.D. degree in computer science and engineering with Delhi Technological University, India. He is also a Lecturer with the Department of Computer Science, Debre Tabor University. His research interests include pattern recognition, deep learning, artificial intelligence, and image processing. He is also dedicated to advancing knowledge and contributing to the development of innovative solutions in these areas.