A Survey on Bangla OCR Processing and Pre-procesing Methodologies

***Abstract*—The necessity of Optical Character Recognition in today's world is a must. It helps many sectors to automate the works, saving times and reduce the chances of errors severely. Researches are working hard to make the OCR system more accurate and better day by day. And Bangla OCR has become a essential research area because of high demand of the language and the numbers of the user. So this paper is all about the comparison of the recent researches of Bangla OCR by their result of accuracy. This research also portraits close encounter of methodologies proposed by some previous researchers. According to the findings, the major factors for getting accuracy is the quality of dataset and the model that researchers implemented. This paper also discussed about the future scopes and possible improvement of Bangla OCR. The report concludes by addressing the potential utilisation in present date.**

***Index Terms*—BPNN; CNN; LSTM; Optical Character Recognition.**

I. Introduction

The technology is improving day by and having a great impact in our daily life. Artificial intelligence is one the advanced technologies. For many decades Artificial Neural Network is being used for character recognition [9]. Optical character recognition is playing a key role here. Many researchers are working on the field of Natural Language Processing but most of them are actually developing systems for language like English, Chinese and Japanese [2]. From 80's developers started working on Bangla OCR and nowadays became a major research area because of it's necessity [11]. But Bangla OCR but it need more attention and it's reuired more and more attention because it is 7th most popular among all languages which is spoken by 250 million peoples all over the world [12]. And the language is rich in terms of history and scripts. Majority of Bangla speaking people lives in India and Bangladesh. [12]. Bangla OCR has an great impact on Banking sectors like automation of data entry in banks and paper scanning an lot of activities are done by OCR nowadays. Bangla OCR can be also utilized by using on postal services to read and process informations from the envlops. Already the technology is being used on CCTVs to recognize and keep tracks of cars by taking data from number plates. Students and teachers are also using Bangla OCR to convert a handwritten document to digital document. It increases the efficiency of working and make the things easy to do.

In this paper I compared the methodologies used by the researchers of Bangla OCR and also compared them side by side. And tried to figure out the best method for Bangla OCR and the unique methods approached by researchers.

II. Background

There are several researches on the topic. for example Shamim Ahmed et al. have proposed a unique approach in "Enhancing the Character Segmentation Accuracy of Bangla OCR using BPNN" [2]. They did image acquisition by scanning image and converted it into a binary image. Then they used Otsu's algorithm to detect object and background. After that they got 89.3% accuracy by doing background removal, Noise reduction, Skew Detection and Correction and segmentation.

Farjana Yeasmin Omee et al. recommanded to use some methods as Global Fixed Threshold, Otsu Global Algorithm, Niblack's Algorithm, Adaptive Niblack's Algorithm, Sauvola's Algorithm to convert images into binary images; it's called binarization [11]. After that they used Noise detection, Skew detection, Page Layout analysis(using RLSA, RAST) and character segmentation. Finally they did feature extraction and implemented Artificial Neural Network to do the research work.

Md Zahangir Alom et al. approach Handwritten Bangla Character Recognition (HBCR) using deep neural networks, including Deep Belief Network (DBN), Convolutional Neural Networks (CNN), and variations with features like dropout, Gaussian filters, and Gabor filters, improves recognition of shapes [4]. The proposed method achieves 98.78% recognition on CMATERdb 3:1:1 database. Diverse handwriting styles and scripts present challenges in character recognition, especially for Bangla with its intricate characters. This study pioneers Handwritten Bangla Digit Recognition (HBDR) using deep learning, introducing a novel CNN-Gabor filters-Dropout integration for enhanced results and comprehensive comparison of five approaches. After around fifteen iteration we have reached almost the maximum accuracy. Lastly, they conducted a comparison between their proposed deep learning method (CNN + Gabor + Dropout) and the current leading techniques. These include MLP (Basu et al., 2005), Modular Principal Component Analysis with Quad Tree based Longest-Run (MPCA+QTLR) (Das et al., 2012a), Genetic Algorithm (GA) (Das et al., 2012b), Simulated Annealing (SA) (Das et al., 2012b), and Sparse Representation Classifier (SRC) (Khan et al., 2014) algorithms for Handwritten Bangla Digit Recognition (HBDR) using the same database. The recognition performance of these approaches is provided for comparison.

AKM Shahariar Azad Rabby et al. used 3 datasets named BanglaLekha-Isolated, CMATERdb and ISI. And they got accuracy of 95.71%, 98%, 96.81% respectively [12]. They

preproccessed the dataset. They proposed a model with 13-layer convolutional neural network and 2 sub-layers that uses ADAM optimizer. Then followed the processes Data augmentation, Training the model, Evaluate the model to get result.

Asif Isthiaq et al. implemented Machine Learning framework TensorFlow for Bangla OCR [9]. They created Tensons, appointed operations bwtween those Tensors and initialized them then created and run session. They made their own dataset for the research and they got result of 71.23% accuracy using by MLP and got 68.82% accuracy by implementing NN.

Md. Akkas Ali et al. approach utilizing a Backpropagation Feed-forward neural network to identify characters based on shape analysis and distinctive features. The determination of optimal hidden layer node count for maximizing the network's performance in recognizing handwritten Bangla characters [3]. The process involves several steps: first, the creation of a binary image; second, the extraction of relevant features to construct an input vector; finally, the application of this input vector within the neural network. Through experimentation, the proposed method achieves an 84% accuracy rate while maintaining lower computational costs compared to alternative techniques.

Angshul Majumdar propesed a on his research using Digital Curvelet Transform and K-NN created a model which is able to the accuracy of 98.60% overall accuracy [10]. He used K-NN for feature extraction and used Fast Fourier Transform algorithm. Moreover he used Sub-band Decomposition, Smooth Partitioning, Renormalization and Morphological Thinning and Thickening.

Md. Hadiuzzaman Bappy et al. approach to the challenge of recognizing handwritten Bengali numerals involves diverse applications such as OCR, postal code identification, and bank check processing [5]. The significance of accurate identification within documents has been acknowledged. With ten numeral classes, the uniqueness of individual writing styles complicates differentiation-using datasets. Achieving accuracy with the complex NumtaDB dataset is challenging, while cleaner datasets like CMATERDB and ISI offer solutions that are more straightforward. NumtaDB combines six datasets with augmented data. Effective feature acquisition and algorithm selection, including SVM, CapsNet, CNN, Logistic Regression, Decision Trees, and KNN, are crucial. The proposed deep CNN model performs well, aided by a two-step preprocessing method. This approach involves image manipulation and augmentation for improved accuracy in classifying transformed images.

Md Abul et al. recommend new OCR for Bangla script recognized as tesseract, by integrating the Tesseract recognition as a script processing power of Bangla OCR. In this research the author built the combined OCR by implementing strategy [8]. In Tesseract OCR the algorithm is used in different stages as the English alphabet which is implemented as a new script. Here is a graphical implementation like by loading a text image to recognize the image then check the spelling error to generate suggestions for error words which can improve the accuracy. The author did his research by using the Tesseract engine. In research firstly preparing training data like(basic, vowel modifiers, consonant modifiers, compound character). Thus it prepared different sets of training data like font type and size, image DPI information, type of document image, segmentation and degradation. Then preprocessing the document image, here the purpose is to collect the information of character units(position of left right and top bottom). In preparing the Tesseract supported image is to store the image until the recognition output text gets. Tesseract engine's goal is to recognize the image and got the output in text. This Tesseract based on Bangla OCR application is affordable for windows and linux environments.

Md Mahadi Hasan Nahid et al. approach Bangla Hand Written Text Recognition- Segmentation based Approach to recognize Bangla numeral using Deep CNN [1]. This study introduces a novel system to recognize continuous handwritten Bangla numerals. A unique Contour Tracing Algorithm (CTA) is employed to segment individual digits from input images, followed by a Convolutional Neural Network (CNN) for numeral recognition. CNN utilizes deep learning to extract features and patterns, simplifying the process of feature engineering. The architecture of CNN includes Convolution, MaxPooling, and Fully Connected layers, facilitating accurate recognition by learning from extensive data. This study presents an innovative segmentation-based approach that introduces a segmentation algorithm (CTA) and a CNN-based recognition system. This method focuses on recognizing entire numeric words instead of single digits, showcasing strong word recognition and improved character recognition and segmentation accuracy. While the segmentation algorithm effectively segments digits within continuous Bangla numerals, its scope is restricted to digits and doesn't cover segmenting handwritten Bangla words with alphabets. With further refinement, the algorithm could potentially enable the recognition of handwritten Bangla words.

PARTHA SARATHI MUKHERJEE et al. in the initial phase of the proposed OCR system involves extracting foreground text from images with noisy backgrounds [6]. A prominent challenge in degraded document OCR is handling characters that are broken, faint, or partially missing in the text. Traditional global thresholding methods struggle with uneven illumination, while local techniques based on heuristics lack overall recognition accuracy for degraded documents. Most recent text extraction techniques focus on complex background texture but overlook recovering broken characters. Though auto encoders have been used for text segmentation, their drawbacks and the need for annotated training data limit their practicality, particularly for low-resource languages like Bangla. Instead, a Generative Adversarial Network (GAN) based segmentation model is employed due to its advantages. To address the lack of training data, a novel Markov Random Field (MRF) based foreground segmentation model is introduced, utilizing unlabeled data for improved appearance modeling. The proposed strategy involves two steps: first, extending the MRF model to extract text from documents

using statistical priors, and second, refining segmentation with a GAN module that incorporates MRF results. This two-step approach enhances segmentation performance and reduces the requirement for pixel-level annotations in document images

Nadim Mahmud et al. proposed "Bangla OCR using Deep Learning based on Image classification Algorithm" suggested three different convolutional neural networks based on the basis of image classification models which are trained and also examined on the BanglaLekhaIsolated dataset [7]. In this research the author firstly collects BanglaLekha Isolated dataset like various character image transformation techniques for preprocessing the data. Here using preprocessing and augmentation, Inception V3, VGG16, Vision Transformer method. The purpose is to recognise Bangla OCR from images. Among three deep learning models VGG16 finds the highest rate of 98.93 percent. Moreover the model focuses on the construction of a state-of-the-art, which is acceptable in Bangla OCR system. By utilizing pre-trained deep learning models, it has implemented a transfer learning technique.

Farisa et al. proposed another complete OCR that shows an end to end OCR system that recognizes Bangla words from images [13]. It can be implemented based on end to end architecture and it is based on Bangla writing. The architecture is based on four different pre-trained CNN architectures and uses two different bidirectional RNN. For improving the system a neural network is required rather than a pre-trained network. From word images by using a method can recognize handwritten Bengali words. Firstly Data Preprocessing then features extraction where CNN architectures, bidirectional RNN model basically used to assume the word from the images. LSTM and GRU used to remove the gradient problems. Then find a loss and error calculator.

III. Methodologies

The researchers worked on different methodologies to build their model and getting more accuracy. Some of the methods are described below:

A. BPNN

Classification and prediction tasks are usually done by using Backpropagation Neural Network. is a type of artificial neural network. In a BPNN, the neural network is trained by feeding it input data and adjusting the weights of the connections between neurons based on the error between the actual output and the desired output. The backpropagation algorithm is used to propagate the error back through the network, adjusting the weights at each layer to reduce the error. BPNNs are popular because they can learn complex nonlinear relationships between inputs and outputs, and can generalize well to new data.

B. CNN

Image processing and video processing tasks are usually done by using Convolutional Neural Network. This is a type of artificial neural network. In a CNN there are multiple layers including convolutional layers, pooling layers and fully connected layers. The convolutional layers works as filter for extracting features from input images. And the pooling layers down sample the images. Afterall connected layers are used to classify image from the extracted layers. With CNN it is possible to handle complex and high dimensional data.

C. LSTM

Long Short-Term Memory is a subset of RNN. It's usually used for sequential data processing. It's being used for speech and text data processing. In this method network there's a memory cell that can remember specific information or forget depending on input signal. The cell plays the major role here. It helps LSTM to store information for long period of time and avoids vanishing gradient problem that happens it terms of classic RNN. It also has the capability to control the information flow in it and out of the memory cell and helps to perform different tasks and inputs. The method LSTM is used in Speech Recognition, Natural Language Processing and many fields nowadays.

D. RLSA

RLSA stands for Run-Length Smoothing Algorithm, which is a technique used in image processing and OCR to improve the quality of binary images. The RLSA algorithm works by scanning a binary image horizontally or vertically and creating a new binary image where each pixel value corresponds to the length of the continuous line of white pixels in the original image. This new image is then thresholded to produce a smoothed binary image where the noise and gaps in the original binary image have been reduced. RLSA is commonly used in OCR systems to preprocess scanned documents before applying character recognition algorithms, as it can help to improve the accuracy and speed of character recognition by reducing noise and enhancing the contrast between characters and background.

IV. Figures and Tables

a) Comparing the accuracy: After comparing all the researches a side-by-side data comparison is given below:

V. Discussion

The present paper provides a comprehensive survey on existing OCR Bangla methodologies including binarization, segmentation, classification, feature extraction and also post-processing. One of the most challenging part on Bangla OCR is character segmentation as it has most complex scripts. That's why horizontal segmentation, vertical segmentation, matra detection are needed. Future research should also explore the use of hybrid OCR techniques that combine different segmentation, feature extraction, and classification methods to achieve better recognition performance.

In this research "Tesseract based on Bangla OCR" the accuracy mostly depends on quality of input image and image resolution. Large amount of dataset experiments to find out the right combination. We observed high and low accuracy of the document image types. In clean printed document type

TABLE I
ACCURACY COMPARISON

Table Number	Details		
	Authors	***Dataset***	***Accuracy***
1	Shamim Ahmed et al.	N/A	89.3%
2	AKM Shahariar et al.	Isolated	95.71%
3	AKM Shahariar et al.	CMATER	98%
4	AKM Shahariar et al.	ISI	96.81%
5	Asif Isthiaq et al.	Own Dataset	71.23%
6	Angshul Majumdar	N/A	96.80%
7	Abul Hasnat et al.	N/A	93%
8	N. M. Dipu et al.	N/A	98.65%
9	Md Zahangir et al.	Own Dataset	98.78%
10	Md. Akkas Ali et al.	Own Dataset	84%
11	Md. Hadiuzzaman et al.	Own Dataset	96.02%
12	PARTHA SARATHI et al.	Own Dataset	81.66%
13	Md Mahadi Hasan et al.	BanglaLekha-Isolated	98.89%
14	Md Mahadi Hasan et al.	Ekush-male	96.15%
15	Md Mahadi Hasan et al.	Ekush-female	96.27%

[a]collected from research papers

we find the accuracy rate 93%. On the other hand printed books & newspapers find the accuracy rate 85%. Moreover, screen print images find the accuracy rate 70% which is the lowest accuracy rate. Clean printed documents have the highest accuracy rate.

In this research the result of accuracy where in model InceptionsV3 training accuracy is 98.77% and test accuracy is 97.82%. On the other hand, the VGG-16 model where training accuracy is 99.23% and test accuracy is 98.65%. Lastly Vision Transformer model where training accuracy is 97.56% and test accuracy is 96.88%. By training the dataset the result of VGG-16 accuracy rate is high. In comparison the worst result of the Model is the vision transformer where the accuracy result is lowest.

REVIEW FINDINGS

From table it's visible that the dataset named CMATER is working good with the model created with Convolutional Neural Network. It has 98% accuracy.

On the other hand, the approach of Shamim Ahmed et al. is unique as they did "Matra Removal" for getting more accurate result.

Every researchers has their won methods and they also have advantage and disadvantages. For future work I think if the "Matra Removal" method is added to the work of preprocessing of CMATER data-set it'll give more accurate result.

CONCLUSION

In conclusion, utilizing the power of Artificial Intelligence to make the work easy it a must. Working on Bangla OCR will help to detect the cars number plate better, scanning of any documentation will be lot easier, Govt. offices and National ID related works can be automated by this technology. The better performance of OCR system required a large data-set and a good model to perform well and enhancing accuracy.

REFERENCES

[1] Md Ahmed, Sk Uddin, and Md Mahadi Hasan Nahid. Bangla handwritten text recognition: A segmentation-based approach to recognize bangla numeral using deep cnn. 01 2023.

[2] Shamim Ahmed and Kashem M.A. Enhancing the character segmentation accuracy of bangla ocr using bpnn. *International Journal of Science and Research (IJSR)*, 2:157–161, 12 2013.

[3] Shamim Ahmed, Nazmus Sakib, Ishtiaque Mahmud, Samiur Rahman, and Md Belali. The anatomy of bangla ocr system for printed texts using back propagation neural network. *Global Journal of Computer Science and Technology*, 12:29–38, 03 2012.

[4] Md. Zahangir Alom, Paheding Sidike, Tarek Taha, and Vijayan Asari. Handwritten bangla digit recognition using deep learning. 05 2017.

[5] Md. Hadiuzzaman Bappy, Md. Siamul Haq, and Kamrul Talukder. Bangla handwritten numeral recognition using deep convolutional neural network. *Khulna University Studies*, pages 863–877, 11 2022.

[6] Ayan Chaudhury, Partha Mukherjee, Sudip Das, Chandan Biswas, and Ujjwal Bhattacharya. A deep ocr for degraded bangla documents. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 21, 04 2022.

[7] Nadim Mahmud Dipu, Sifatul Alam Shohan, and K. M. A. Salam. Bangla optical character recognition (ocr) using deep learning based image classification algorithms. In *2021 24th International Conference on Computer and Information Technology (ICCIT)*, pages 1–5, 2021.

[8] Md Hasnat, Muttakinur Chowdhury, and Mumit Khan. An open source tesseract based optical character recognizer for bangla script. pages 671–675, 01 2009.

[9] Asif Isthiaq and Najoa Saif. Ocr for printed bangla characters using neural network. *International Journal of Modern Education and Computer Science*, 12:19–29, 04 2020.

[10] Angshul Majumdar. Bangla basic character recognition using digital curvelet transform. *Journal of Pattern Recognition Research*, 1:17–26, 01 2007.

[11] Farjana Yeasmin Omee, Shiam Shabbir Himel, and Md. Abu Naser Bikas. A complete workflow for development of bangla ocr, 2012.

[12] Akm Shahariar Azad Rabby, Sadeka Haque, Sanzidul Islam, Sheikh Abujar, and Syed Akhter Hossain. Bornonet: Bangla handwritten characters recognition using convolutional neural network. *Procedia Computer Science*, 143:528–535, 2018. 8th International Conference on Advances in Computing Communications (ICACC-2018).

[13] Farisa Benta Safir, Abu Quwsar Ohi, M. F. Mridha, Muhammad Mostafa Monowar, and Md. Abdul Hamid. End-to-end optical character recognition for bengali handwritten words, 2021.