Bangla handwritten word recognition using YOLO V5

Md. Anwar Hossain[1], AFM Zainul Abadin[1], Md. Omar Faruk[1], Iffat Ara[1], Mirza AFM Rashidul Hasan[2], Nafiul Fatta[1], Md Asraful[1], Ebrahim Hossen[1]

[1]Department of Information and Communication Engineering, Faculty of Engineering and Technology, Pabna University of Science and Technology, Pabna, Bangladesh

[2]Department of Information and Communication Engineering, Faculty of Engineering, University of Rajshahi, Rajshahi, Bangladesh

Article Info

Article history:

Received Jun 11, 2023
Revised Aug 18, 2023
Accepted Oct 24, 2023

Keywords:

Bengali word detection
Cursive handwriting
Handwritten documents
Recurrent neural networks
You only look once

ABSTRACT

This research paper presents an innovative solution for offline handwritten word recognition in Bengali, a prominent Indic language. The complexities of this script, particularly in cursive writing, often lead to overlapping characters and segmentation challenges. Conventional methodologies, reliant on individual character recognition and aggregation, are error-prone. To overcome these limitations, we propose a novel method treating the entire document as a coherent entity and utilizing the efficient you only look once (YOLO) model for word extraction. In our approach, we view individual words as distinct objects and employ the YOLO model for supervised learning, transforming object detection into a regression problematic to predict spatially detached bounding boxes and class possibilities. Rigorous training results in outstanding performance, with remarkable box_loss of 0.014, obj_loss of 0.14, and class_loss of 0.009. Furthermore, the achieved mAP_0.5 score of 0.95 and map_0.5:0.95 score of 0.97 demonstrates the model's exceptional accuracy in detecting and recognizing handwritten words. To evaluate our method comprehensively, we introduce the Omor-Ekush dataset, a meticulously curated collection of 21,300 handwritten words from 150 participants, featuring 141 words per document. Our pioneering YOLO-based approach, combined with the curated Omor-Ekush dataset, represents a significant advancement in handwritten word recognition in Bengali.

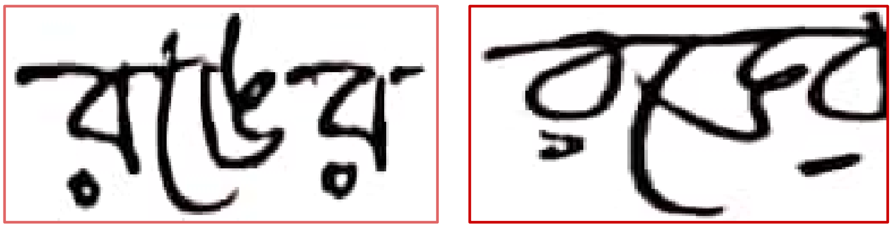
Corresponding Author:

Md. Anwar Hossain
Department of Information and Communication Engineering, Faculty of Engineering and Technology
Pabna University of Science and Technology
Rajapur, Pabna, Bangladesh
Email: manwar.ice@gmail.com

1. INTRODUCTION

There are around 300 million Indians globally, and Bengali is their primary language of communication. Bengali is the official language of Bangladesh and the Republic of India. It is among the most extensively used writing systems worldwide used by over 265 million people [1]. Word recognition is a technique that enables computers to identify written or printed words and convert them into a format that the computer can understand. Bangla word identification in handwritten writing is one of the utmost fascinating areas of research in today's world, and it is becoming more and more popular. Even in the twenty-first century, handwritten communication has its place and is virtually always used in daily life as a manner of capturing information that is intended to be shared with others. The need for online information systems has grown along with the expansion of the internet. Bangla handwritten word recognition research encompasses a wide range of applications, making it a vital field of study. These applications include document digitization, post offices, banks, document analysis and recognition, education, and the preservation of historical documents, signature verification, and other institutions [1]. It also plays a crucial role in enhancing accessibility tools, streamlining e-commerce processes, and facilitating signature verification. Furthermore, the integration of handwriting recognition in personal assistants, note-taking apps, and search engines significantly improves user experience and caters to the needs of millions of Bengali speakers worldwide. The identification of Bangla handwritten words holds paramount significance, demanding substantial focus to advance various active applications in the field. Word detection in Bangla handwritten text is challenging due to complex alphabet shapes, variations caused by cursive writing, uneven lighting, and image distortions. Traditional techniques struggle with ambiguity, noise, and lack of standardization [2]. Innovative approaches and advanced models are necessary to improve word recognition for efficient document processing and applications in education and communication. To overcome the limitations in Bengali handwritten word recognition, we propose an innovative approach that treats the entire document as a coherent entity. By leveraging the efficiency of the you only look once (YOLO) model, we precisely extract words by viewing them as distinct objects. Through supervised learning, we transform object recognition into a regression problem, enabling us to predict spatially detached bounding containers and class possibilities with accuracy. Rigorous training of our YOLO model leads to outstanding performance, resulting in precise and efficient word recognition in Bengali sentences. Additionally, we curate a comprehensive and unique Bengali dataset named Omor Ekush containing complex words, and make it openly available for future research and advancements. This pioneering research significantly pushes the boundaries of Bengali handwritten word recognition, opening up new avenues for improved document processing and analysis capabilities. Extracting the words from scanned images containing handwriting in the Bengali language is the main objective of this paper.

Figure 1(a) illustrates a sample of comparatively good handwriting in Bangla. The writing is clear, well-formed, and easily legible. In disparity, Figure 1(b) displays an example of comparatively cursive handwriting in Bangla, where characters are conjoined and the writing style is more fluid. Cursive handwriting poses challenges in word recognition due to overlapping and tilted characters, making it difficult for conventional recognition techniques.



(a) (b)

Figure 1. Comparatively handwriting; (a) comparatively good handwriting and (b) comparatively cursive handwriting

Figure 2 showcases the process of word recognition from handwritten Bangla text. The demo illustrates how the proposed method effectively identifies and extracts individual words from the handwritten text. By treating the entire document as a coherent entity and using the YOLO model, the system accurately predicts spatially separated bounding boxes and class probabilities for each word. Rigorous training ensures outstanding performance, demonstrating the model's exceptional accuracy in recognizing and extracting words from handwritten Bangla text.

Figure 2. Word recognition from handwriting

2. RELATED WORKS

In this segment, we provide an overview of the existing studies in the arena of Bangla handwritten word appreciation. Previous research efforts in this area have primarily revolved around the acknowledgment of handwritten Bangla words. The central focus of prior investigations has been on developing methodologies and algorithms capable of accurately identifying and interpreting complete words written in the Bengali script. This task is particularly challenging due to the cursive nature of the calligraphy, which often leads to significant variations in character shapes and connectivity within a word. By examining the completed tasks related to Bangla handwritten word recognition, we gain insights into the progress made in this domain and the potential areas for further advancement. As we delve into the literature, we aim to highlight the gaps and opportunities that exist for developing more robust and sophisticated word recognition systems in the context of Bengali handwriting.

Roy *et al.* [3] made the initial breakthrough in Bengali word recognition while working on Indian postal automation. They concentrated on recognizing Bengali words that consisted of exactly 76 city names and employed the 2D non-symmetric half plane-hidden Markov model (HMM) method for accomplishing this task. A more comprehensive analysis of their effort can be initiate in [4].

Pal *et al.* [5] proposed a model that they focused on unrestrained handwritten town title detection, utilizing a lexicon-driven tactic. They implemented the water-reservoir technique to segment city word images into primitives and then utilized dynamic programming to achieve finest character separation. In conclusion, they adopted a controlling element constructed modified quadratic discriminant function categorizer for the possibility calculation of the letterings.

Bhowmik *et al.* [6] utilized a HMM to identify calligraphic city titles with the aid of a fixed-size lexicon. The HMM was trained using genetic algorithms. Their approach included a structural feature that employed a directional encoding scheme on boundaries. In a subsequent study [7], the same group made further improvements by dipping the lexicon dimensions, effectively narrowing down the search interplanetary. This reduction was achieved through an analysis of the perpendicular and straight hits of the word image. These advancements aimed to enhance the accuracy and efficiency of handwritten town name recognition, making notable contributions to the field of handwriting recognition research.

Roy *et al.* [8] anticipated an HMM founded Bengali handwritten term recognizer. In their recognition module, they employed zone-wise horizontal dissection surveyed by vertical separation in the central region. They utilized the local gradient histogram as a feature set and fed it to a left-to-right HMM for recognizing the mid zone. A gradient feature-based support vector machine (SVM) classifier was employed to recognize the upper and lower zone modifiers. Next, they combined the zone-wise predictable results via a character arrangement approach.

Bhunia *et al.* [9] utilized five types of features for middle zone recognition and compared their approach with their earlier work in [10]. They further combined their findings from [8], [9] to develop a unified approach for Devanagari and Bengali word acknowledgement in [11]. The experimental dataset used across these studies included various word images, incorporating numerous city names. Their cohesive approach contributes significantly to the advancement of word recognition in Bengali script and provides valuable insights for similar applications in other scripts.

Adak *et al.* [12] introduced a neural network (NN) grounded system for Bengali handwritten word detection, via convolutional neural network (CNN) with long short-term memory (LSTM). They curated a novel dataset called NewISIdb also employed a unified architecture, combining CNNs with a recurrent model. LSTM blocks and connectionist temporal classification (CTC) layer further enhanced the recognition accuracy, showcasing the potential of neural nets in this domain. The study represents a significant advancement in Bengali handwritten word appreciation research.

In our research, we employed an innovative solution for handwritten word detection in Bengali. Our novel method utilizes the YOLO model to process the entire document as a coherent entity, accurately predicting bounding boxes and class probabilities for word extraction. Rigorous training leads to outstanding performance, demonstrating exceptional accuracy in detecting and recognizing handwritten words. The curated Omor-Ekush dataset enhances our approach, providing a diverse benchmark for word recognition models. This pioneering YOLO-based method signifies a significant advancement in handwritten word recognition in Bengali, with transformative potential in document analysis, automated transcription, and language processing.

Table 1 provides a comprehensive summary of research endeavors in the field of handwritten Bengali word recognition, highlighting the approaches employed and the research gaps addressed by each study. The table offers valuable insights into the contributions of each study and the specific areas in the literature they have targeted for improvement.

Table 1. Overview of handwritten Bengali word recognition studies and research gaps

Study	Focus	Approach	Research limitations
Roy *et al.* [3]	Bengali word recognition	2D NSHP-HMM technique for recognizing 76 city names	Limited research on recognizing specific Bengali words, especially city names
Pal *et al.* [5]	Handwritten city name recognition	Lexicon-driven approach, water-reservoir technique, and MQDF classifier	Scarcity of methods for recognizing unconstrained handwritten city names in Bengali script
Bhowmik *et al.* [6]	Handwritten town name recognition	HMM with genetic algorithms, structural feature	Need for efficient techniques to identify handwritten city names, particularly with lexicon of fixed size
Roy *et al.* [8]	HMM-based Bengali Handwritten word recognizer	Region-wise horizontal and vertical segmentation, HMM, SVM classifier, character alignment strategy	Advancements in zone-wise recognition for Bengali handwritten words
Bhunia *et al.* [9]	Middle zone recognition	Contrast of features, unified approach for Bengali and Devanagari word acknowledgment	Enhancements in recognizing middle zones of Bengali handwritten words
Adak *et al.* [12]	Bengali handwritten word recognition	CNNs with LSTM, YOLO model, NewISIdb dataset, CTC layer	Utilization of CNNs with LSTM for Bengali handwritten word recognition, the introduction of the NewISIdb dataset
Proposed method	Offline handwritten word recognition in Bengali	YOLO model to treat entire documents, predict bounding boxes, and class probabilities for word extraction	Addressing complexities in Bengali cursive writing, introducing a novel approach using the YOLO model and Omor-Ekush dataset for offline handwritten word recognition

3. BACKGROUND STUDY

The most recent task demonstrates that using artificial neural networks is advantageous for maintaining the image detection task. In the 1950s, researchers first developed concepts such as the perceptron learning algorithm [13]. Recent NNs base their operations on theories developed during the perceptron period. This unit begins by defining a neuron as a crucial component of contemporary NNs. Then it goes into further detail on CNNs and recurrent neural networks (RNNs).

3.1. Perceptron

The perceptron, an artificial neuron, forms the fundamental building block of modern neural networks, pivotal in machine learning. Thus, a perceptron is combined with a sole neuron. The activation a, which is the yield of the neuron, is mapped by Φ: RN → R as (1):

$$\alpha = \Phi(x) = \sigma(wTx + b) \tag{1}$$

The weights $w \in R^N$ and $b \in R$ and the nonlinear function σ (·) map the feature vector $x \in R^N$. On the occasion of the perceptron σ (·) viewpoints for:

$$\sigma(z) = \{1 \quad if\ z > 0\ \ otherwise\ 0 \tag{2}$$

3.2. Convolutional neural networks

Convolution networks which has obligated a significant impression on the arena of image analysis and has been the basis for many recent advances in deep learning [14]. A CNN is a NN that processes an image also represents with a vector code. CNNs are founded upon fully connected neural networks in their architecture. Likewise, a convolutional network comprises multiple layers that process signals and propagate them forward. Nevertheless, unlike a route activation found in a fully allied layer, CNN activations possess the structure of three-dimensional tensors. Typically termed a “feature map,” this output tensor plays a crucial role in CNN. An example of this transformation is when the first convolutional layer takes an input image with dimensions $3 \times W \times H$ and produces a feature map with dimensions $H' \times W' \times C$. Here, C represents the quantity of features extracted by the layer. Essentially, a convolutional layer converts one volume into another. A standard CNN is composed of multiple convolutional stage, with dense layers at the top that transform the final convolutional dimensions obsessed by a vector output. In technical terms, the vector representation of an image is commonly referred to as fc7 features, as the seventh fully connected layer of the alexnet architecture originally provided the source of acquisition for this data [15]. Despite the fact that many newer architectures have surpassed the performance of alexnet, and current state-of-the-art designs differ from it, the term “fc7 features” has remained popular within the field. Moreover, one can include an extra layer, like a soft-max layer, atop the fc7 features, depending on the particular problem that the network aims to address. A typical CNN architecture is illustrated in Figure 3.

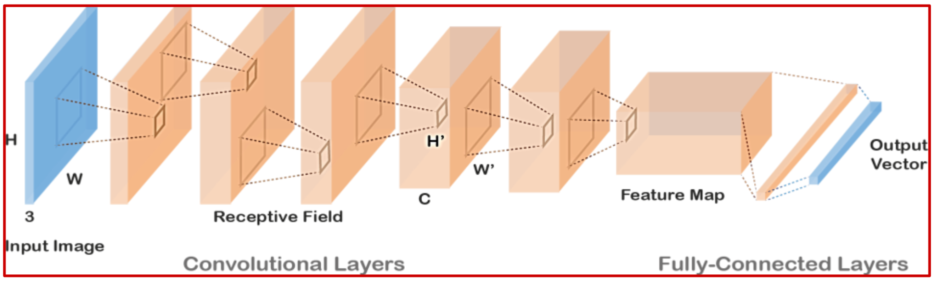


Figure 3. Layer with CNN

3.3. Pooling layer

The design of convolutional layers allows for the preservation of spatial dimensions while increasing the depth of the network as information flows through it. However, the thing is applied to diminish three-dimensional, specifically in advanced layers. Dimensions diminish may be gained by exercising tread after convoluting, top to a series of interested areas join. Nevertheless, an extra forthright procedure was established termed a pooling layer. The input is divided and hooked on non-overlapping regions. And the layer produces a grid containing the extreme values from each region. Pooling layers are commonly used amongst convolutional layers in order to decrease the dimensionality of the data.

3.4. Weights

In a NN, each neuron produces an output by applying a specific function to the inputs it receives from the receptive field of the preceding layer. A weight vector and a bias determine the purpose which is pragmatic to the input data. Iteratively fine-tuning these biases and weights is what learning is all about. Filters are vectors of weights and biases that represent specific structures of the input. A unique characteristic of CNNs is the ability for multiple neurons to utilize the same filter. The use of a shared filter, with a single bias and weight vector, across multiple receptive fields reduces the memory footprint of the network, compared to having a separate bias and weight vector for each receptive field.

4. DATASETS

4.1. Previous datasets

Handwriting acknowledgment has been a captivating research issue for almost fifty years. While early successes were achieved in recognizing simple handwritten digits. In 1992, the first census of optical character recognition system conference organized the pioneering large-scale character recognition challenge [16]. Following this, researchers gradually began to develop datasets for offline handwriting recognition at the sentence and document levels for the English language [17], [18]. The identity and access management (IAM) dataset was later utilized to launch one of the most well-known shared tasks in handwriting recognition the International Conference on Document Analysis and Recognition (ICDAR) challenge [19]. Our research revealed only a limited number of Bangla calligraphy datasets, with the widely held consisting of compound character data. BanglaLekha-Isolated [20] is an example of such a dataset. It comprises 10 numerals, 50 simple characters, and 24 cautiously chosen multifarious characters. The dataset includes 2000 individual images for each of the 84-character classes. After removing any scribbles, the final dataset contains a total of 166,105 pictures of handwritten fonts. The dataset additionally incorporates information regarding the age and gender of the individuals who contributed the handwriting samples.

Research by Rabby *et al.* [21] contains an alternative versatile handwritten character dataset that also contains 367,018 characters. The data was gathered from various regions of Bangladesh, with an equivalent number of female and male contributors from a range of age groups. In addition to the character data, the dataset also includes a set of modifiers, a feature not found in other similar character-level datasets. In addition to other resources, the ISI [22] and CMATERdb [23] datasets represent dual of the earliest collections of handwritten characters for the Bangla linguistic. The Bangla writing [10] dataset is the sole collection that bears similarity to our dataset with respect to word-level annotation. The dataset encompasses the handwriting samples from 260 individuals, varying in age and character. The creators utilized an explanation instrument to label the piece of paper with bounding boxes that enclose the unicode depiction of the words. This collection encompasses a significant number of 32,787 characters and 21,234 words, boasting a vocabulary size of 5,470. Despite the fact that all word label bounding boxes were created physically, the actual pounded truth of the pages from which the texts were generated was not provided. Furthermore, the majority of the pages were brief and could be perceived as resembling a paragraph rather than a complete document.

4.2. Dataset preparation

This article introduces a Bangla handwriting dataset called Omor Ekush, which includes single-page handwriting samples from 150 individuals of varying ages and characters. Every page contains bounding boxes that enclose a piece of word, as well as the unicode depiction of the text. This collection encompasses 21,300 words. All bounding boxes were initially created using our custom segmentation script, then manually verified and labeled. The dataset is suitable for advanced optical character/word acknowledgment, author proof of identity, handwritten word separation, and word generation.

4.3. Dataset description

Omor Ekush the dataset introduced in this article, strives to offer a superior handwriting collection that is enhanced in all aspects. The data set can be utilized for a range of claims based on machine learning and deep learning techniques. The dataset is applicable for writing biometric tasks, including identification and proof. Moreover, this technology shows promise in specific computer vision applications, including recognizing optical characters and segmenting handwriting. In addition, the dataset possesses the potential to power procreant calligraphy models. The construction and utilization of this dataset differ from typical Bangla datasets [20]. Current datasets for Bangla script only include individual character examples. In contrast, the Omor Ekush dataset includes word-based script with bounding boxes, similar to [10]. The dataset was built using established offline handwriting and writer recognition datasets as a foundation [18]. The majority of larger datasets, such as KHATT [24] and IAM [18], incorporate automated and pre-determined parameters for data labeling. In contrast, the Omor Ekush dataset's annotations and labels were manually created.

4.3. Dataset generation

4.3.1. Data acquisition

We want our dataset to cover all the characters in Bengali language, that's why we carefully chose two Bengali pangrams from the wikipedia. Figure 4 depicts the selected pangrams. We break the line spacing in the verse pangram to save the page space. As the rhythm is not important for the dataset. Combining both pangrams, a single document contains approximately 141 words, but due to some word duplication the number of unique words becomes 137. Figure 5 represents the sample data taken from write then according to perform our model.

"বর্ষামুখর দিন শেষে, উর্দ্ধপানে চেয়ে যখন আষাঢ়ে গল্প শোনাতে বসে ওসমান ভুঁইঞা, ঈষান কোণে তখন অন্ধকার মেঘের আড়ম্বর, সবুজে ঋদ্ধ বনভূমির নির্জনতা চিরে থেকে থেকে ঐরাবতের ডাক, মাটির উপর শুকনো পাতা ঝরে পড়ে ঔদাসীন্যে, এবং তারই ফাঁকে জমে থাকা ঢের পুরোনো গভীর দুঃখ হঠাৎ যেন বৃষ্টিতে ধুয়ে মুছে ধূসর জীবনে রঙধনু এনে দেয়।"

"হৃদয়ের চঞ্চলতা বন্ধে ব্রতী হলে জীবন পরিপূর্ণ হবে নানা রঙের ফুলে। কুজ্ঝটিকা প্রভঞ্জন শঙ্কার কারণ লণ্ডভণ্ড করে যায় ধরার অঙ্গন। ক্ষিপ্ত হলে সাঙ্গ হবে বিজ্ঞজনে বলে শান্ত হলে এ ব্রহ্মাণ্ডে বাঞ্ছিতফল মেলে। আষাঢ়ে ঈশান কোনে হঠাৎ ঝড় উঠে গগন মেঘেতে ঢাকে বৃষ্টি নামে মাঠে ঊষার আকাশে নামে সন্ধ্যার ছায়া ঐ দেখো থেমে গেছে পারাপারে খেয়া। শরৎ ঋতুতে চাঁদ আলোয় অংশুমান সুখ দুঃখ পাশাপাশি সহ অবস্থান। যে জলেতে ঈশ্বর তৃষ্ণা মেটায় সেই জলেতে জীবকুলে বিনাশ ঘটায়। রোগ যদি দেহ ছেড়ে মনে গিয়ে ধরে ঔষধের সাধ্য কী বা তারে সুস্থ করে?"

Figure 4. Selected Bengali pangrams

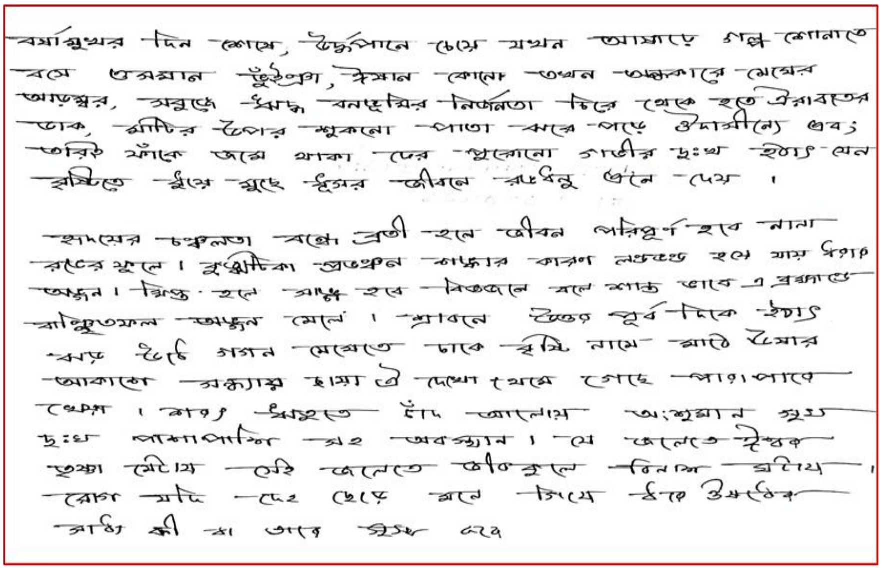


Figure 5. Sample image data taken from writer

The dataset was gathered from students at Pabna University of Science and Technology. All the students are between 20-25 both male and female. The authors used A4-sized paper and a consistent ball-point pen for script. Participants received instructions to write about a pre-selected topic. Thus, individual text comprises approximately the same quantity of words.

From Figure 6 we can see that five of the classes have a higher frequency than the rest of the classes. This is because each pangram has common words in them. Table 2 includes the common words of each pangram.

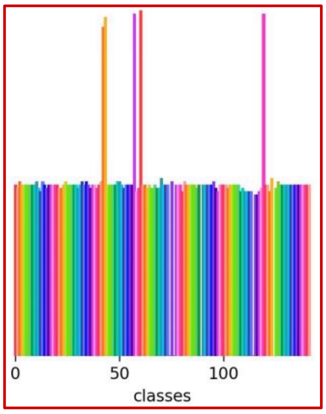


Figure 6. Words (classes) distribution of the dataset

Table 2. Common words of each pangram

Label	Class1	Class2
হলে	74	57
হলে	60	76
হঠাৎ	43	89
দুঃখ	42	114
জলেতে	119	124

4.3.2. Data extraction

The handwritten images are digitized using smartphone cameras with 1280p resolution. The dataset encompasses a total of 150 images, captured using smartphone cameras. These pictures are processed using CamScanner App to remove all the unwanted noise from the image. For example, various lighting effects, flashlight glares, and shadow effects. But in some images, still some noise remains.

4.3.3. Data processing and labeling

Processing and labeling huge amount of data is trivial task. Initially our approach was to extract bounding region of words based on the space between the words in a sentence with the help of OpenCV find contours method (which finds the all-connected components of an image). But due to cursive handwriting style words gets smashed to each other. So automated script is not an ideal option for us and had to do it manually. For manual annotation software like LabelMe [25] each word's bounding box needs to be hand drawn even though we can auto generate bounding boxes for some word based on the spacing exploit. That's why we created a completely new annotation software handwritten automated word annotation (HAWAN).

5. HANDWRITTEN AUTOMATED WORD ANNOTATION

HAWAN is a web-based software to ease the handwritten word annotation including bounding box generation and labeling. It's generated annotation info into a JSON format.

Figure 7 show the initial prediction of the word bounding box based on the OpenCV find contours method. Initial prediction contains lots of unwanted bounding boxes. HAWAN provides some sophisticated actions to remove all unwanted bounding boxes. It's also providing drag and drop word labeling to ease the annotation task.

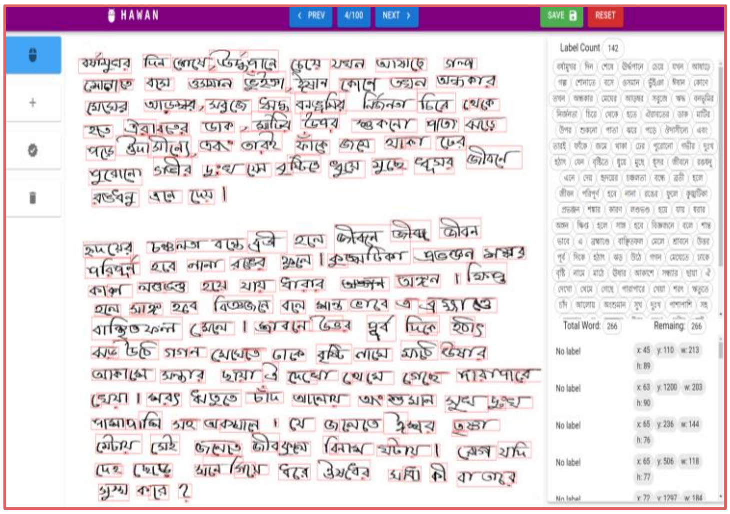


Figure 7. HAWAN

Figure 8 shows properly annotated words. If the bounding boxes are annotated it turn's bounding boxes color from red to green for visual inspection. Once all the bounding boxes are properly annotated, save action generates a JSON file with all the bounding boxes information and filename information. The bounding-box and label data for individually image were stored in specific JSON files. The specifying agreement adhered to that of the handwritten images. Figure 9 displays the parameters of the standard JSON file. To maintain the authenticity and quality of the dataset, we did not apply any augmentation to increase its size.

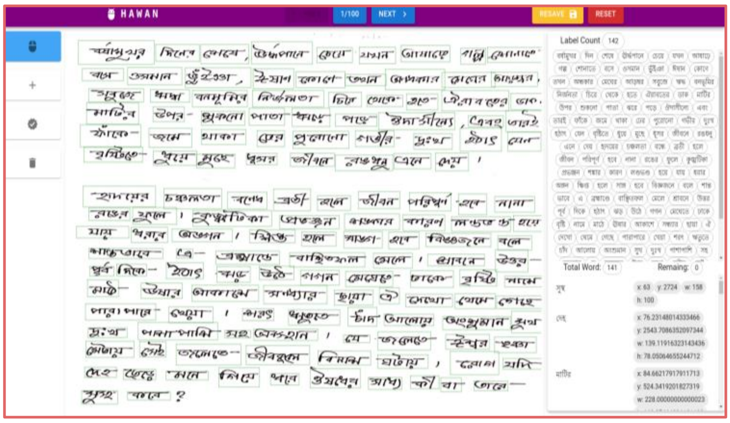


Figure 8. Properly annotated words

```
{
    "boxes": [
        {
            "h": 83.79100031357444,
            "id": 68,
            "label": "\u09a6\u09c1\u0983\u0996",
            "w": 157.71405589752922,
            "x": 198.08888939554657,
            "y": 369.9136119472259
        },
        ...
    ],
    "file": "sharifAhmad190636_2.jpg"
}
```

Figure 9. Annotation information in JSON format

6. YOLO V5 MODEL

YOLO is interestingly modest. A sole convolutional network is exploited to expect numerous bounding boxes and their corresponding class possibilities. YOLO employs training on complete images and straight enhances recognition performance. This integrated model offers some advantages over conventional approaches to object detection. As YOLO v5 is similar to other single-stage object detectors, YOLO, being a single-stage object detector, consists of three crucial components. Backbone, neck, and head, Figure 10 demonstrations the straightforward architecture of YOLO v5.

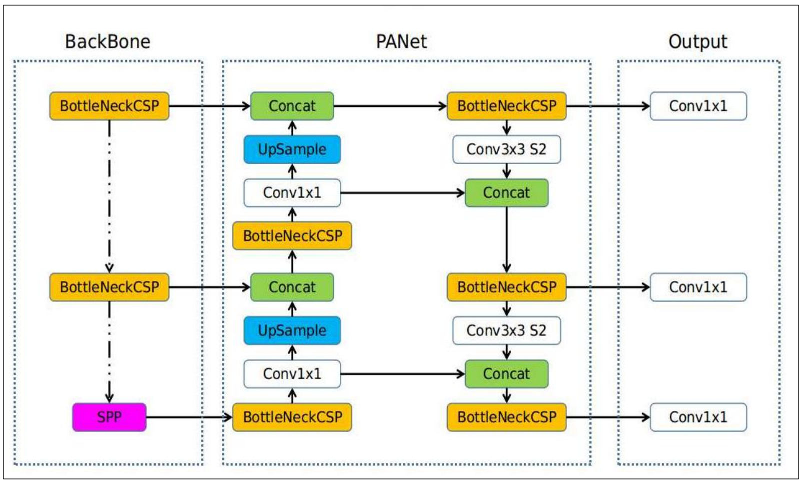


Figure 10. YOLO v5 simplified architecture

In YOLO v5, the Leaky ReLU activation function is employed in the hidden layers, while the sigmoid activation function is utilized in the ultimate detection layer. SGD is the default optimizer function for the YOLO v5. YOLO v5 computes the loss by utilizing a weighted sum of three distinct losses:

a. The class loss in YOLO v5 employs binary cross entropy (BCE);
b. The objectness loss in YOLO v5 utilizes BCE;
c. The location loss in YOLO v5 which is intersection over union (IoU).

IoU loss is calculated by factoring overlap between the prediction and ground truth.

7. RESULT AND DISCUSSION

7.1. Model training

We explored a range of techniques to train the model. Initially, we trained the model using batch sizes of 64 and an image size of 640. But due to low GPU memory issue batch size 6 is selected. We conducted our training with SGD optimizer and with default hyperparameters and anchors. Figure 11 depict the data labeling in training time.

7.2. Metrics

Our object detection models have two responsibilities: to locate the bounding box of an object and to predict the label assigned to that box. Mean average precision is the standard method for assessing the performance of mutually tasks performed by an object recognition model. When evaluating object detection models, we must assess not only the model's classification abilities but also its ability to accurately locate objects. As a result, we require a distinct evaluation metric, known as the mean average precision (mAP). The fundamental concept is to combine the evaluation of detection and classification abilities. The method used by the mean average precision to determine the accuracy of a bounding box is called the IoU. The IoU for a predicted hopping box P, Q and its corresponding minced truth label is computed using (3):

$$IoU(P, Q) = \frac{Area(P \cap Q)}{Area(P \cup Q)} \tag{3}$$

By utilizing IoU values, we investigate various methods for determining true positives and false positives. If the IoU value meets a certain threshold, we classify it as a true positive. For instance, if we establish a threshold of 0.5, an example with an IoU greater than 0.5 is classified as a true positive, while those with an IoU less than or equal to 0.5 are considered false positives. To calculate precision, divide the number of true positives by the total number of positive predictions. This can be expressed as (4):

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (4)$$

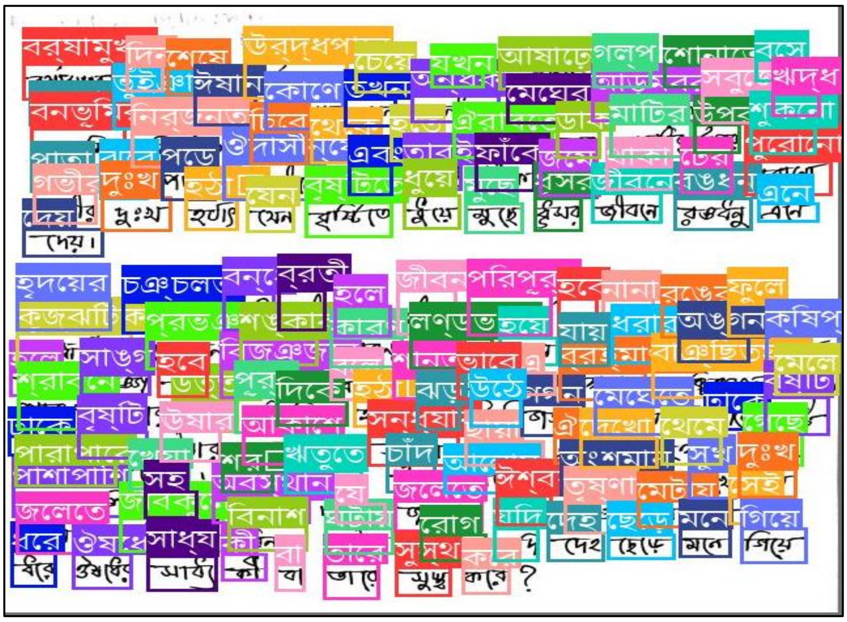


Figure 11. Training data with labeling

We compute metrics for IoU thresholds extending from 0.5 to 0.95 in increments of 0.05. The average precision for a class is then determined by averaging the precision values for that class crossways all IoU thresholds. Finally, the mAP for n classes, where k represents each class, is computed as (5):

$$mAP = \frac{1}{n} \Sigma AP_r \quad (5)$$

The percentage of total predictions that the model properly recognizes is known as accuracy. Recall is the measure of the fraction of actual positives that the model correctly identifies. The F-1 score signifies the harmonic mean of recall and precision. Mathematically, the metrics are specified as (6) to (8):

$$accuracy = \frac{True\ Positives(TP) + True\ Negatives(TN)}{True\ Positives + True\ Negatives + False\ Positives + False\ Negatives} \quad (6)$$

$$recall = \frac{TP}{FP + TP} \quad (7)$$

$$F1 = \frac{2(Precision * Recall)}{Precision + Recall} \quad (8)$$

8. EVALUATION

During the training process, we monitored the act of the primary metrics on the validation set and tracked the loss to ensure that we were not overfitting. Let us first examine the loss graphs, which are divided into CLS loss, objective loss, box loss. Here from Figure 12, we see that ‘box_loss’, ‘obj_loss’ and ‘class_loss’ gets down smoothly till 290 epochs but after 290 epochs obj_loss tends to increase, a sign of overfitting. So, we stopped our training. Table 3 include evaluation time performance of our model.

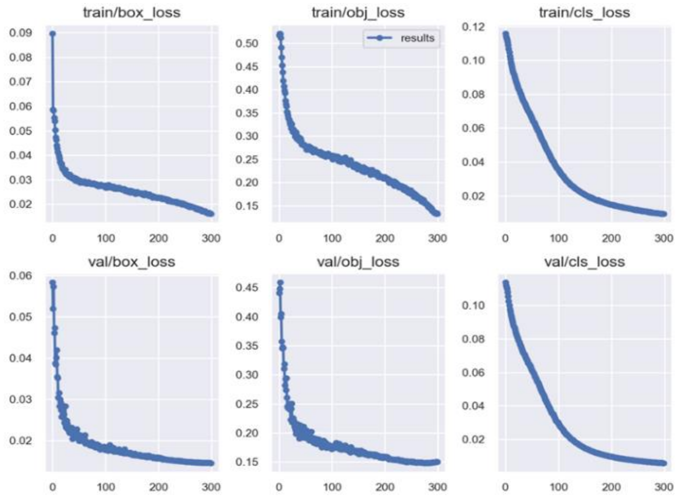


Figure 12. Box_loss, obj_loss and class_loss graph

Table 3. Evaluation time performance metrics

Loss	Train dataset	Val dataset
Box_loss	0.01	0.01
Obj_loss	0.14	0.15
Class_loss	0.009	0.01

During training, we tracked the mAP for the range of 0.5 to 0.95. Figure 13 shows that, over epochs, there is indeed an inverse correlation between the mAP and the loss, as anticipated. From the Figure 14 we can see that machine generated document from the physical handwritten document in Figures 15(a) and (b). We can also see that there are some ambiguities while predicting certain classes (bolded text). This problem can be eliminated by growing the number of samples in the dataset.

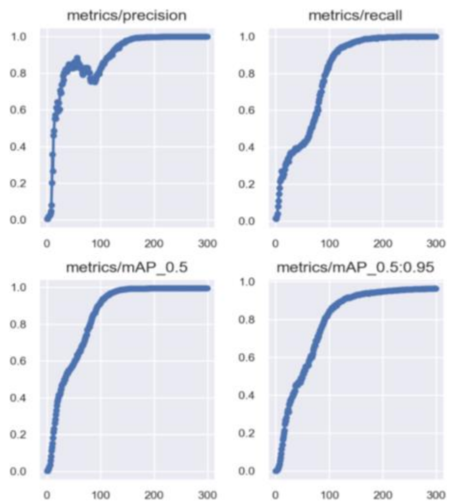


Figure 13. mAP50 and mAP95 graph

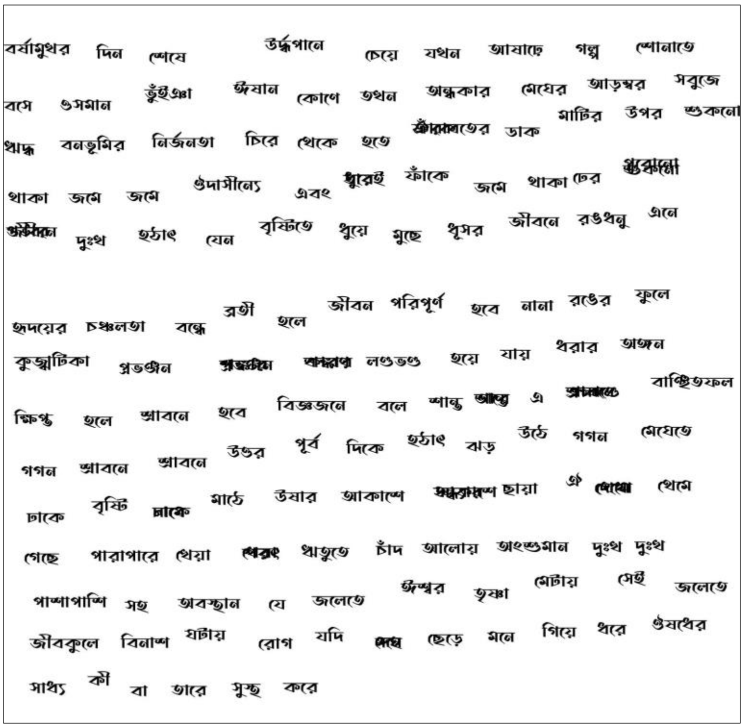
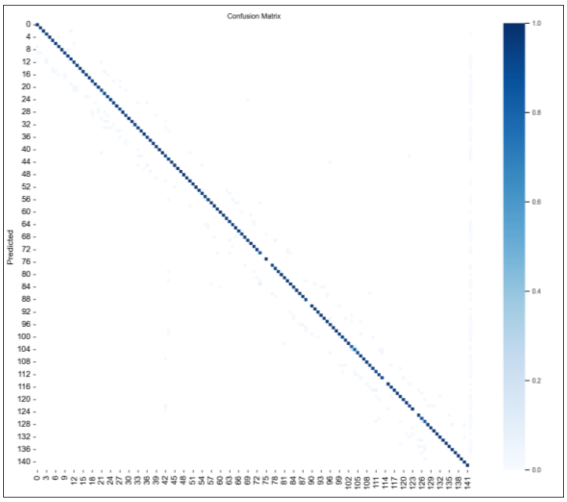


Figure 14. Simplified predicted output

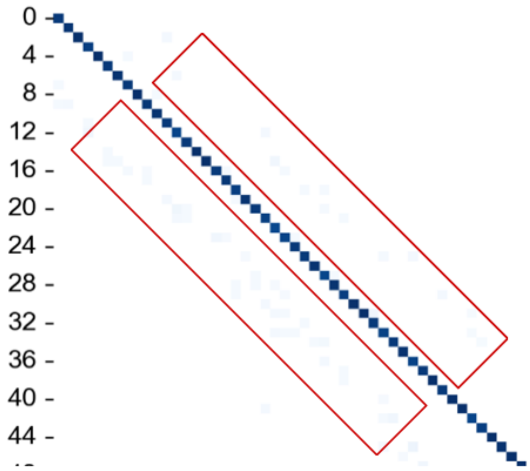
(a) (b)

Figure 15. Handwritten input and output; (a) handwritten input given to the model and (b) predicted output with accuracy

From the confusion matrix in Figure 16(a), it may seem like class 74,76,89,114,124 are not classified properly but class (74,57), (76,60), (89,43), (114,42), (124,119) are same class. Due to lots of class it is printed as a heatmap for better understanding. Figure 16(b) indicates that some classes get misclassified with low probability (low color intensity), which can be eliminated by increasing the number of epochs in the training.



(a)



(b)

Figure 16. Matrix; (a) confusion matrix and (b) confusion matrix (zoomed)

9. CONCLUSION

In our research, we have addressed the task of recognizing Bengali words from handwritten documents. To facilitate research in this domain, we have developed and made publicly available our own dataset named “Omar Ekush”. The inclusion of this dataset will significantly aid future studies in the field of Bengali word recognition. Additionally, we have developed a custom annotation software called HAWAN, which streamlines the process of dataset annotation, further enhancing the accessibility and usability of the dataset. Our experiments have demonstrated the effectiveness of the YOLO v5 model trained on the “Omar Ekush” dataset for word recognition in handwritten images. The model showcased impressive performance, as measured by the mAP metrics. This indicates that our model can accurately and robustly identify Bengali words from handwritten documents. Furthermore, the ‘Omar Ekush’ dataset’s depth and diversity extend its utility beyond word recognition to areas like handwriting analysis and document understanding. Our annotation tool, HAWAN, has potential for adaptation to various languages. Together, they provide a robust foundation for future research, fostering innovation in handwriting recognition and document analysis. In conclusion, our work significantly contributes to academic and practical aspects of Bengali word recognition and handwriting analysis. We encourage researchers and practitioners to utilize the 'Omar Ekush' dataset and explore HAWAN's capabilities to advance handwritten text recognition and related fields. In the future, we aim to improve Bengali word recognition by expanding the dataset with diverse data, introducing more data classes, integrating the model with HAWAN for better annotations, optimizing hyperparameters, and utilizing a more powerful GPU for faster training and experimentation. These efforts will lead to enhanced accuracy and robustness in recognizing Bengali words from handwritten documents.

ACKNOWLEDGEMENTS

We appreciate the help with this research from the Department of Information and Communication Engineering at Pabna University of Science and Technology.

REFERENCES

[1] M. Asraful, M. A. Hossain, and E. Hossen, "Handwritten Bengali Alphabets, Compound Characters and Numerals Recognition Using CNN-based Approach," *Annals of Emerging Technologies in Computing (AETiC)*, vol. 7, no. 3, pp. 60–77, 2023, doi: 10.33166/AETiC.2023.03.003.

[2] Q. Ye and D. Doermann, "Text detection and recognition in imagery: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 7, pp. 1480-1500, 2014, doi: 10.1109/TPAMI.2014.2366765.

[3] K. Roy, S. Vajda, U. Pal, B. B. Chaudhuri, and A. Belaïd, "A system for Indian postal automation," *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, vol. 2, pp. 1060-1064, 2005, doi: 10.1109/ICDAR.2005.259.

[4] S. Vajda, K. Roy, U. Pal, B. B. Chaudhuri, and A. Belaid, "Automation of Indian postal documents written in Bangla and English," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 08, pp. 1599-1632, 2009, doi: 10.1142/S0218001409007776.

[5] U. Pal, K. Roy, and F. Kimura, "A lexicon driven method for unconstrained Bangla handwritten word recognition," in *International Workshop on Frontiers in Handwriting Recognition*, pp. 601-606, 2006.

[6] T. K. Bhowmik, S. K. Parui, and U. Roy, "Discriminative HMM training with GA for handwritten word recognition," in *2008 19th International Conference on Pattern Recognition*, Tampa, FL, USA, pp. 1-4, 2008, doi: 10.1109/ICPR.2008.4761830.

[7] T. K. Bhowmik, U. Roy, and S. K. Parui, "Lexicon reduction technique for Bangla handwritten word recognition," in *2012 10th IAPR International Workshop on Document Analysis Systems*, Gold Coast, QLD, Australia, pp. 195-199, 2012, doi: 10.1109/DAS.2012.50.

[8] P. P. Roy, P. Dey, S. Roy, U. Pal, and F. Kimura, "A novel approach of Bangla handwritten text recognition using HMM," in *2014 14th International Conference on Frontiers in Handwriting Recognition*, Hersonissos, Greece, pp. 661-666, 2014, doi: 10.1109/ICFHR.2014.116.

[9] A. K. Bhunia, A. Das, P. P. Roy, and U. Pal, "A comparative study of features for handwritten Bangla text recognition," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, Tunis, Tunisia, pp. 636-640, 2015, doi: 10.1109/ICDAR.2015.7333839.

[10] M. F. Mridha, A. Q. Ohi, M. A. Ali, M. I. Emon, and M. M. Kabir, "BanglaWriting: A multi-purpose offline Bangla handwriting dataset," *Data in Brief*, vol. 34, p. 106633, 2021, doi: 10.1016/j.dib.2020.106633.

[11] P. P. Roy, A. K. Bhunia, A. Das, P. Dey, and U. Pal, "HMM-based Indic handwritten word recognition using zone segmentation," *Pattern recognition*, vol. 60, pp. 1057-1075, 2016, doi: 10.1016/j.patcog.2016.04.012.

[12] C. Adak, B. B. Chaudhuri, and M. Blumenstein, "Offline cursive Bengali word recognition using CNNs with a recurrent model," in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 429-434, 2016, doi: 10.1109/ICFHR.2016.0086.

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998, doi: 10.1109/5.726791.

[14] M. Otahal, M. Najman, and O. Stepankova, "Design of neuromorphic cognitive module based on hierarchical temporal memory and demonstrated on anomaly detection," *Procedia Computer Science*, vol. 88, pp. 232-238, 2016, doi: 10.1016/j.procs.2016.07.430.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84-90, 2017, doi: 10.1145/3065386.

[16] R. A. Wilkinson *et al.*, "The first census optical character recognition system conference," *US Department of Commerce, National Institute of Standards and Technology*, 1992.

[17] U.-V. Marti and H. Bunke, "A full English sentence database for off-line handwriting recognition," in *Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR '99 (Cat. No.PR00318)*, Bangalore, India, 1999, pp. 705-708, doi: 10.1109/ICDAR.1999.791885.

[18] U.-V. Marti and H. Bunke, "The IAM-database: an English sentence database for offline handwriting recognition," *International Journal on Document Analysis and Recognition*, vol. 5, pp. 39-46, 2002, doi: 10.1007/s100320200071.

[19] M. Zimmermann and H. Bunke, "Automatic segmentation of the IAM off-line database for handwritten English text," in *2002 International Conference on Pattern Recognition*, Quebec City, Canada, 2002, vol. 4, pp. 35-39, doi: 10.1109/ICPR.2002.1047394.

[20] M. Biswas *et al.*, "Banglalekha-isolated: A multi-purpose comprehensive dataset of handwritten bangla isolated characters," *Data in Brief*, vol. 12, pp. 103-107, 2017, doi: 10.1016/j.dib.2017.03.035.

[21] A. S. A. Rabby, S. Haque, M. S. Islam, S. Abujar, and S. A. Hossain, "Ekush: A multipurpose and multitype comprehensive database for online off-line bangla handwritten characters," in *RTIP2R 2018: Recent Trends in Image Processing and Pattern Recognition*, vol. 1037, pp 149–158, 2019, doi: 10.1007/978-981-13-9187-3_14.

[22] U. Bhattacharya and B. B. Chaudhuri, "Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 3, pp. 444-457, 2008, doi: 10.1109/TPAMI.2008.88.

[23] R. Sarkar, N. Das, S. Basu, M. Kundu, M. Nasipuri, and D. K. Basu, "CMATERdb1: a database of unconstrained handwritten Bangla and Bangla–English mixed script document image," *International Journal on Document Analysis and Recognition (IJDAR),* vol. 15, pp. 71-83, 2012, doi: 10.1007/s10032-011-0148-6.

[24] S. A. Mahmoud *et al.*, "Khatt: Arabic Offline Handwritten Text Database," in *2012 International Conference on Frontiers in Handwriting Recognition,* Bari, Italy, 2012, pp. 449-454, doi: 10.1109/ICFHR.2012.224.

[25] A. Torralba, B. C. Russell, and J. Yuen, "Labelme: Online image annotation and applications," *Proceedings of the IEEE,* vol. 98, no. 8, pp. 1467-1484, 2010, doi: 10.1109/JPROC.2010.2050290.

BIOGRAPHIES OF AUTHORS

Md. Anwar Hossain received B.Sc. (Honours) and M.Sc. degrees in Information and Communication Engineering from the University of Rajshahi, Bangladesh in 2005 and 2006 respectively. He received his M.Phil. degree from the Pabna University of Science and Technology, Bangladesh in 2020. In 2010 he served as a lecturer in the Department of Information and Communication Technology of Comilla University, Bangladesh. In 2012, he joined Pabna University of Science and Technology, Bangladesh as a faculty member, where he is currently serving as an Associate Professor in the Department of Information and Communication Engineering. Now, he is a Ph.D. student in the Department of Information and Communication Engineering (ICE), Pabna University of Science and Technology (PUST), Bangladesh. His research interests include deep learning, machine learning, image classification, and natural language processing. He can be contacted at email: manwar.ice@gmail.com.

AFM Zainul Abadin received the B.Sc. and the M.Sc. degrees in Information and Communication Engineering from the University of Rajshahi, Bangladesh, in 2006 and 2007 respectively. He received the M.Phil. degree in wireless communication with the specialization of physical layer security assisted NOMA technology from Pabna University of Science and Technology, Bangladesh in 2021. Currently, he is an Associate Professor of Information and Communication Engineering department, Pabna University of Science and Technology, Bangladesh and a Ph.D. research fellow in the Universiti Kebangsaan Malaysia. His research interests include information security, image steganography, intelligent systems, data science, deep learning, image processing, and computer vision. He can be contacted at email: abadin.7@gmail.com.

Md. Omar Faruk earned his B.Sc. (Hon's), M.Sc., and Ph.D. from the University of Rajshahi, Bangladesh in Applied Physics and Electronic Engineering in 1994, 1996, and 2012, respectively. He worked at Science Workshop, University of Rajshahi, Bangladesh, as an assistant instrument engineer from 2001 to 2004, an instrument engineer from 2004 to 2008, a senior instrument engineer from 2008 to 2011, and a principal instrument engineer from 2011 to 2013. He accepted a position as an assistant professor at the Pabna University of Science and Technology in Pabna, Bangladesh, in the Department of Information and Communication Engineering in 2013. In 2019, he received a promotion to Associate Professor. Seismology, machine learning, and the internet of things (IoT) are some of his areas of interest. He can be contacted at email: fom_06@yahoo.com.

Iffat Ara was born in 1986 in Pabna, Bangladesh. She graduated from Rajshahi University in Bangladesh in 2010 with a Master's degree and a B.Sc. (Honors) in Applied Physics and Electronic Engineering. At Pabna University of Science and Technology in Bangladesh, she is now an Associate Professor in the Department of Information and Communication Engineering. Her study focuses on the evaluation of bio-medical signals. She can be contacted at email: ara.iffat@ymail.com.

Mirza AFM Rashidul Hasan graduated from the University of Rajshahi in Bangladesh with a B.Sc. (Hons), an M.Sc., and an M.Phil. in Applied Physics and Electronic Engineering in 1992, 1993, and 2001, respectively. He began teaching at the University of Rajshahi in Bangladesh in 2006, and he presently holds the position of Professor in the Department of Information and Communication Engineering. From 2003 to 2004, he served as a guest researcher at Waseda University in Japan. From 2006 to 2007, he served as a junior fellow at IWMI. He earned his Ph.D. from the Islamic University's Faculty of Applied Science and Technology in Kushtia, Bangladesh, in 2009, and his D.Eng. from the Graduate School of Science and Engineering at Saitama University in Saitama, Japan, in 2012. His current areas of interest are voice, picture, and communication systems applications of digital signal processing. He can be contacted at email: mirzahasanice@gmail.com.

Nafiul Fatta received the B.Sc. (Engg.) degree from the Department of Information and Communication Engineering (ICE), Pabna University of Science and Technology, Pabna, Bangladesh in 2020. His research interests are artificial intelligence, machine learning, deep learning, and natural language processing. He can be contacted at email: nafiul.ice.pust@gmail.com.

Md Asraful is an ICE student who is currently studying in the Faculty of Engineering at Pabna University of Science and Technology, Pabna. He has chosen a Bachelor of Information and Communication Engineering (ICE). His research interests are artificial intelligence, machine learning, deep learning, and natural language processing. He can be contacted at email: mdasrafulm333@gmail.com.

Ebrahim Hossen is an ICE student who is currently studying in the Faculty of Engineering at Pabna University of Science and Technology, Pabna. He has chosen a Bachelor of Information and Communication Engineering (ICE). His research interests are artificial intelligence, machine learning, deep learning, and natural language processing. He can be contacted at email: ebrahim.180611@s.pust.ac.bd.