Bangla Handwritten Compound Character Recognition

BY

Md. Al - Jubair

151-15-4691

Sayeed Tanzim

151-15-4813

Jyoti Chandra

151-15-4735

This Report Presented in Partial Fulfillment of the Requirements for B.Sc in CSE (Computer Science & Engineering)

Supervised By

Dr. Sheak Rashed Haider Noori

Associate Professor and Associate Head

Department of CSE, Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
November 2018

APPROVAL

This Project/internship titled **“Bangla Handwritten Compound Character Dataset”**, submitted by **Md. Al-jubair**, **ID No: 151-15-4691** and **Sayeed Tanzim**, **ID No: 151-15-4813 and Jyoti Chandra, ID No: 151-15-4735** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on November 25, 2018.

BOARD OF EXAMINERS

Dr. Syed Akhter Hossain **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

MD. Tarek Habib **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mr. Narayan Ranjan Chakraborty **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Mohammad Shorif Uddin **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that this research has been done by us under the supervision of **Dr. Sheak Rashed Haider Noori,** Associate Professor and Associate Head, Department of CSE, Daffodil International University. We also declare that neither this research nor any part of this research has been submitted elsewhere for the award of any degree.

SUPERVISED BY:

Dr. Sheak Rashed Haider Noori
Associate Professor and Associate Head
Department of CSE
Daffodil International University

SUBMITTED BY:

Md. Al-Jubair
ID: 151-15-4691
Department of CSE
Daffodil International University

Sayeed Tanzim
ID: 151-15-4813
Department of CSE
Daffodil International University

Jyoti Chandra
ID: 151-15-4735
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing that made us possible to complete the final year project successfully.

We really grateful and wish our profound indebtedness to **Dr. Sheak Rashed Haider Noori**, Associate Professor and Associate Head, Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Machine Learning*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to the Almighty Allah and Head**,** Department of CSE, for his kind help to finish our project and also to other faculty members and the staff of CSE department of Daffodil International University.

We would like to thank our all course mates in Daffodil International University, who took part in the discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Developing new algorithms or findings related to any language requires a rich dataset. In Bangla language, most of the datasets are consist of 50 basic characters and a few compound characters. Though many researchers have worked on Bangla printed and handwritten characters but still, the dataset of most compound characters (handwritten) is not done yet. In this paper, a dataset of 76 handwritten compound characters and 3 modifiers are pre-processed. All those characters are unique and absent in other datasets of Bangla characters. The dataset is available publicly at [https://bit.ly/2PBEIrj].

TABLE OF CONTENTS

CONTENTS **PAGE**

Board of Examiners…………………………………………………………......... i

Declaration………………………………………………………………………… ii

Acknowledgment……………………………………………………………... iii

Abstract…………………………………………………………………………. iv

CHAPTER **PAGE**

CHAPTER 1: INTRODUCTION 11-13

1.1 Introduction………………………………………………………………………… 11

1.2 Motivation……………………………………………………………………...... 11

1.3 Rationale of the Study……………………………………………………….. 11-12

1.4 Research Questions………………………………………………………... 12

1.5 Expected Outcome…..………………………………………………………... 13

1.6 Layout of the Report……………………………………………………….. 13

CHAPTER 2: BACKGROUND STUDY 14-19

2.1 Introduction……………………………………………................................. 14

2.2 Related Works ………………………………..……………………….. 15-16

2.3 Research Summary ……………………..……………………………… 16-18

2.4 Scope of the Problem…………………………………………………………...... 19

2.4 Challenges………………………………………………………………………… 19

CHAPTER 3: RESEARCH METHODOLOGY 20-24

3.1 Introduction…………..…………………………………………………………. 20

3.2 Research Subject and Instrumentation…………………………………….. 20

3.3 Data Collection Procedure…………………………………………………. 20-21

3.4 Proposed Methodology………………………………………………………… 22-25

3.5 Statistical Analysis…………………………………………………….... 25-26

CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSIONS 25-27

4.1 Introduction ………………….……………………………………………… 27

4.2 Experimental Results ……………………………………………………… 27-29

4.3 Descriptive Analysis ………………………………………………………. 29

4.4 Summary ……………………………………………………………………... 29

CHAPTER 5: CONCLUSION, LIMITAIONS & FUTURE WORKS 30

5.1 Summary of the Study.. 30

5.2 Conclusion………………………………………………………………………. 30

5.3 Recommendations………………………………………………………………. 30

5.4 Limitations & Future Works…………………….……………………………. 31

REFERENCES……………………………………………………… 32-34

Plagiarism Report…………………………………………………… 35

LIST OF TABLES

TABLES **PAGE**

Table 2.1 Number of images in different datasets …………………………………… 14

Table 2.2 Overview of Datasets………………………………………………………… 17-18

LIST OF FIGURES

FIGURES **PAGE**

Figure 3.1 Sample of filled data collection form.……………………………….. 21

Figure 3.2 Steps of the proposed methodology……………………………… 22

Figure 3.3 Proposed architecture……………………………………………… 24

Figure 3.4 Gender distribution of contributors ………………………………….. 25

Figure 3.5 Age distribution of contributors ………………………….................. 26

Figure 4.1 Gray scale image with garbage ……………….………….................. 27

Figure 4.2 Gray scale image ……………….……….................................... 28

Figure 4.3 Distribution of gray scale image to individual class ………………….. 28

Figure 4.4 Processed images from dataset …………………………………..... 29

CHAPTER 1
INTRODUCTION

1.1 Introduction

Bangla is the language of Bangladeshis and the seventh language as most spoken language in the world. More than 300 million people speak Bengali and it is the official language of Bangladesh. Bangla is used in different Indian states such as West Bengal, Tripura, Assam and Jharkhand. In this fast-moving technology era, the Bangla handwritten character recognition is a significant challenge which is to be overcome. At the same time, the English language for machine learning tool has achieved great success [5]. Currently, several datasets are easily available for the basic characters of Bangla [9]-[11]. But there is no such dataset available for Bengali compound characters. As a result, the researchers could not do much about the Bangla compound characters [11].

1.2 Motivation

- **Computerize all previous papers:** There are so many previous papers of law related or deeds which is not available all the time we've to blow wound so many papers to find any paper we need. We can make an easy documentation of them to find easily at any time anywhere.
- **Reserving Literature:** There are many writings of writers which have no documentation. Through this project we can make a documentation of that handwritten papers.
- **Incomplete Bangla OCR:** Over 300 million people speak in Bangla but yet Bangla OCR is incomplete.

1.3 Rationale of the Study

Rationale of the study should be specific to understand. It is also important to explain the research ideally. It is important to relate with the following points:

1.3.1 Contribution to the elimination of a gap in the Bangla OCR

The research needs to contribute to the elimination of a gap in the Bangla OCR. Elimination of a gap in existing research is our main target. There is a huge amount of gap in Bangla OCR as it is incomplete. But there is complete OCR in English, Arabic etc.

1.3.2 Conduction to solve a specific problem

Also there is so many works on Bangla OCR on Bangla handwritten normal characters but there is no work on Bangla handwritten compound characters. So it is a specific problem for developing Bangla OCR. For these purpose, huge amount of dataset is needed and it is made manually.

1.3.3 Contribution to the level of professional development of the researcher

These research has a great level of value in professional development of the researchers. There is so many works on these subject but that work is successful which contributes in professional development of the researcher. It is a great platform to gather more knowledge on Bangla language.

1.4 Research Questions

Questions that was tried to answer through this research are:

- Incomplete Bangla OCR
- Incomplete Bangla Handwritten Compound Character Dataset
- Dataset pre-processing technique
- Appropriate/suitable machine learning model to train the dataset

1.5 Expected Outcome

We are going to make the largest dataset of Bangla handwritten compound character. This dataset will be trained by machine learning algorithm.

.

1.6 Layout of the Report

- Chapter one demonstrates an introduction to the project with its motivation, research questions, and expected outcome.
- Chapter two discusses related works, scope of the problem and challenges.
- Chapter three contains research methodology.
- Chapter four covers experimental results and some relevant discussions.
- Chapter five draws a conclusion and discusses limitations and future works.

CHAPTER 2
Background Study

2.1 Introduction

Different languages have various types of datasets in the world for use in OCR. According to the language and type of dataset, there are differences in the size of the dataset and dataset creation process. Among the important languages of the Indian subcontinent, there are datasets in Bengali, Hindi, Tamil, Chinese, Urdu etc. language. Other than this subcontinent, English and Arabic language datasets are notable. Almost every given language has separate alphabetic and numerical datasets.

Bangla52 Dataset is the largest dataset of compound characters in Bangla language that is suitable for use in machine learning algorithms. In addition to this, the demographic data of the individual contributor (age, gender) can be found out of the computational data. The dataset comprises 76 isolated handwritten compound characters and 3 modifiers. Initially, 932 handwriting samples for each of the characters were collected and preprocessed. After reducing glitch and noises 72,512 handwritten character images were found which are convenient to fit into machine learning tools. So, these images were included in the processed dataset. With the unique 932 handwriting samples, there had been collected demographic information such as age, gender and educational labels of the contributors which can be used in different studies [1]-[27].

Table 2.1. Number of images in different datasets

Dataset	**Compound Characters**
CMATERDB[9]	42,248
ISI Dataset[10]	None
BanglaLekha-Isolated[11]	47,407
Bangla52	72,512

Bangla52 Dataset is consist of 76 frequently used compound characters along with 72,512 square images. On the other hand, BanglaLekha-Isolated Dataset contains 24 compound characters. ISI Dataset didn’t involve any compound characters but 30,966 images of Bangla basic characters.

2.2 Related Works

Currently the biggest dataset for the English language ‘NIST Special Dataset 19' of 810000 individual images from 3600 contributors [17]. With the help of ‘NIST Special Dataset 19' dataset, many other datasets have been made available. Among them, A-Z Handwritten alphabets and MNIST database are remarkable [17]. A-Z Handwritten Alphabets Dataset has 3700000+ images of 26 alphabets. On the other hand, there are 60000 and 10000 training and test data respectively in the MNIST dataset. Each image of both the datasets is 28 × 28 pixels in size.

HP Labs India Indic Handwriting Datasets is an immense dataset of three Indian languages [16]. They have made datasets of isolated characters of Tamil, Telugu and Hindi languages, and the dataset of isolated words of Hindi and Tamil. Where there are 500 samples for each of the 156 letters of Tamil language, which have been collected from local Tamil people of India. Such as school students, students of the university and older people. The dataset of Telugu language contains 44820 images of 166 characters. For every single character of 111 Hindi alphabets, there are 270 set samples from the local 100 Hindi speakers. There are 100 set samples for each of the 85 isolated words of Tamil and 220 sets for each of the 70 isolated Hindi words.

UCOM offline dataset is a dataset of Urdu language. This dataset is consist of handwritten 600 pages in Urdu, which is in Nasta'liq Style [15]. CASIA Online and Offline Chinese Handwriting Databases is a prodigious dataset of Chinese language that comprises 3.9 million samples of 7185 Chinese alphabet and 171 symbols [14]. Also, Chinese handwritten text dataset is of 5090 pages that contain 1.35 million alphabets. The Arabic language consists of 28 alphabets. The Arabic handwritten character dataset comprises a total of 28000 images of 28 characters [13]. Also, Arabic handwritten digit database contains 70,000 digits written by 700 contributors [12].

There are several datasets available in Bangla, like all other languages. Among them CMATERDB [9], ISI Dataset [10], BanglaLekha-Isolated [11] are notable. There are 50 basic characters and 300+ compound characters in Bangla language. In CMATERDB, the number of images of basic characters, numerals and compound characters is 15103, 6000 and 42,248 respectively. On the other hand, ISI Dataset contains images of 30,966 basic characters and 23,299 numerals. And the latest BanglaLekha-Isolated comprises 98,950 basic characters, 19,748 Numerals, and 47,407 Compound Character images.

2.3 Research Summary

Our dataset, the Bangla52 consists of 76 Bangla handwriting compound characters and 3 modifiers. The selected compound characters that we chose are frequently used. There have been done remarkable works on Bangla character recognition but these 76 handwritten compound characters are yet to be work on [1] – [27]. The main focus of this research was to make a dataset of these 76 compound characters and 3 modifiers that can be used for further research. Primarily, the raw data was collected on forms from 932 individuals including both male and female of various ages ranging between 10 and 17. Before preprocessing, data with clear mistakes were discarded. The dataset contains a total number of 72,512 digitized images of Bangla handwritten compound characters. The following information is shown in Table 2.2.

Table 2.2 Overview of Datasets

Language	Number of Characters	Name of Datasets	Size of Dataset
Bangla	Basic Character- 50 Compound Character – 300+ Bangla Digit - 10	Bangla52 [this work]	Compound Character – 72,512
		CMATERDB [9]	Basic character – 15103 Compound Character- 42,248 Numeral - 6000
		ISI Dataset [10]	Basic character – 30,966 Compound Character- None Numeral - 23299
		BanglaLekha-Isolated [11]	Basic character – 98950 Compound Character- 47407 Numeral - 19748
English	Basic Character – 26	NIST Special Dataset 19 [17]	Basic character-810000
		A-Z Handwritten alphabets	Basic character-

	English Digit - 10	[17]	3700000
		MNIST [17]	Numeral- 70000
Hindi / Devanagari	Basic character - 111	HP Labs India Indic Handwriting Datasets [16]	Character – 29970 Word – 15400
Tamil	Character - 156		Character – 78000 Word - 8500
Telugu	Character - 166		Character - 44820
Urdu	Letter – 58 Basic character - 39	UCOM offline dataset [15]	600 pages of handwriting
Chinese	Character – 7185 Symbol - 171	CASIA Online and Offline Chinese Handwriting Databases [14]	Isolated character – 3.9 million Text character – 1.35 million
Arabic	Alphabet - 28	Arabic handwritten character dataset [13]	Character - 28000
		Arabic handwritten digit database [12]	Numerals - 70000

2.4 Scope of the problem

There are many scope of problem to make the background of the Bangla OCR. Specially to make the data set. There is the biggest data set of English and so many works on English OCR. But the dataset amount is so poor in Bangla OCR.

There can also be problem in data preprocessing. There is many garbage image that we have to filter them. It also needs so many time in data preprocessing. Sometimes usable data is filtered out and we have to find out them and again we have to set them in data set.

2.5 Challenges

- We need to contact with so many people to collect data, also we have to explain them that what actually we want to do in our research, what the importance of our work is.
- There is another challenge we have to face that is data preprocessing problem. We have to make a large dataset and that is why it is a matter of huge amount of time.
- Need high configuration machine

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Introduction

Data processing is a technique which can convert data into usable and desired form. To accomplish this task, an established sequence of operation should be followed which can be either automatic or manual. Most of the data processing is done by using computers and thus done automatically. The processed data can be obtained in different forms such as image, graph, table, vector file, audio, charts or any other desired format depending on the method of data processing used.

3.2 Research Subject and Instrumentation

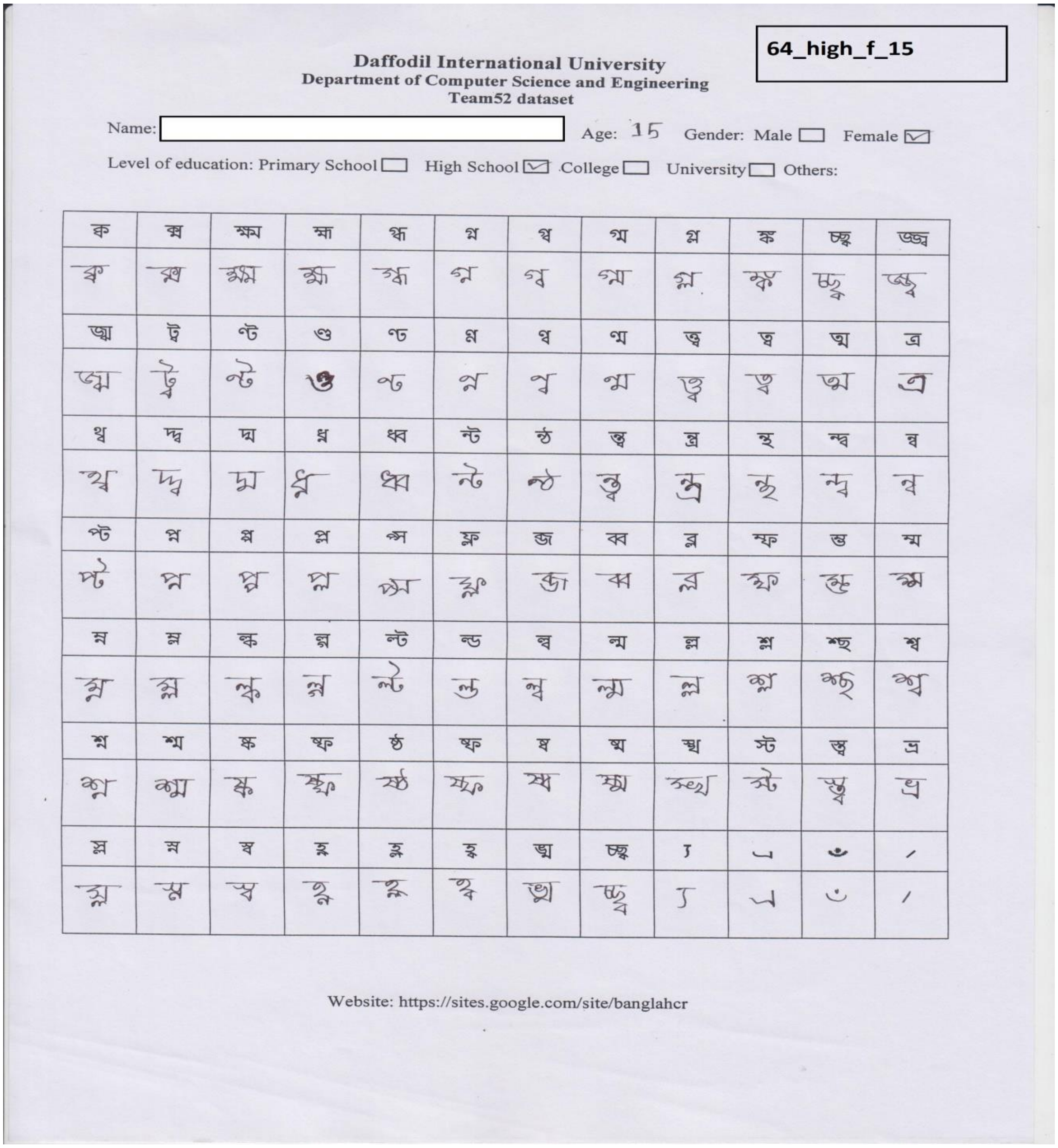
Like all other recognition procedure character recognition is nothing but a recognition process. Lots of recognition system is available in computer science, and also Bangla OCR plays a prominent role in computer science. This report discuss the theory and implementation of an OCR for Bangla. This process is instrumented very carefully with all the process. After completing manual process, we have done our technological process. The work in progress is extending it to handle multiple type of process.

3.3 Data Collection Procedure

A sample of the form that was used to collect handwriting is shown at Fig.3.1. Each individual was asked to fill the form in their own pace.

Every single form was named using both alphabet and numeric numbers. Those names indicate specific information about that form. The first digit of the name means the serial number and the next 4 alphabets indicate the level of education. After that, a single character is used to identify the gender of the subject (f: Female, m: Male). And the last 2 digit shows the age of the contributor. All those four parts of the identification name are separated by underscores (_) [8] [11]. For example:

64_high_f_15



64_high_f_15

Daffodil International University
Department of Computer Science and Engineering
Team52 dataset

Name: Age: 15 Gender: Male □ Female ☑

Level of education: Primary School □ High School ☑ College □ University □ Others:

ক্ক	ক্ম	ক্ষ্ম	ক্ষ্ণ	গ্ধ	গ্ন	গ্ব	গ্ম	গ্ল	ঙ্ক	চ্ছ্ব	জ্জ্ব
ক্ক	ক্ম	ক্ষ্ম	ক্ষ্ণ	গ্ধ	গ্ন	গ্ব	গ্ম	গ্ল	ঙ্ক	চ্ছ্ব	জ্জ্ব
জ্ঝ	ট্ব	ণ্ট	ণ্ড	ণ্ঠ	ণ্ন	ণ্ব	ণ্ম	ত্ত্ব	ত্ব	ত্ম	ত্র
জ্ঝ	ট্ব	ণ্ট	ণ্ড	ণ্ঠ	ণ্ন	ণ্ব	ণ্ম	ত্ত্ব	ত্ব	ত্ম	ত্র
থ্ব	দ্দ্ব	দ্ম	ধ্ন	ধ্ব	ন্ট	ন্ঠ	ন্ত্ব	ন্ত্র	ন্থ	ন্দ্ব	ন্ব
থ্ব	দ্দ্ব	দ্ম	ধ্ন	ধ্ব	ন্ট	ন্ঠ	ন্ত্ব	ন্ত্র	ন্থ	ন্দ্ব	ন্ব
প্ট	প্ন	প্প	প্ল	প্স	ফ্ল	ব্জ	ব্ব	ব্ল	ম্ফ	ম্ভ	ম্ম
প্ট	প্ন	প্প	প্ল	প্স	ফ্ল	ব্জ	ব্ব	ব্ল	ম্ফ	ম্ভ	ম্ম
ম্ন	ম্ল	ল্ক	ল্গ	ল্ট	ল্ড	ল্ব	ল্ম	ল্ল	শ্ল	শ্ছ	শ্ব
ম্ন	ম্ল	ল্ক	ল্গ	ল্ট	ল্ড	ল্ব	ল্ম	ল্ল	শ্ল	শ্ছ	শ্ব
শ্ন	শ্ম	ষ্ক	ষ্ফ	ষ্ঠ	ষ্ফ	ষ্ব	ষ্ম	স্খ	স্ট	স্ত্ব	ভ্র
শ্ন	শ্ম	ষ্ক	ষ্ফ	ষ্ঠ	ষ্ফ	ষ্ব	ষ্ম	স্খ	স্ট	স্ত্ব	ভ্র
স্ল	স্ন	স্ব	হ্ন	হ্ল	হ্ব	ভ্ম	চ্ছ্ব	্য	্র	ঁ	্
স্ল	স্ন	স্ব	হ্ন	হ্ল	হ্ব	ভ্ম	চ্ছ্ব	্য	্র	ঁ	্

Website: https://sites.google.com/site/banglahcr

Fig. 3.1 Sample of filled data collection form.

3.4 Proposed Methodology

This is how our proposed methodology works for preprocessing:

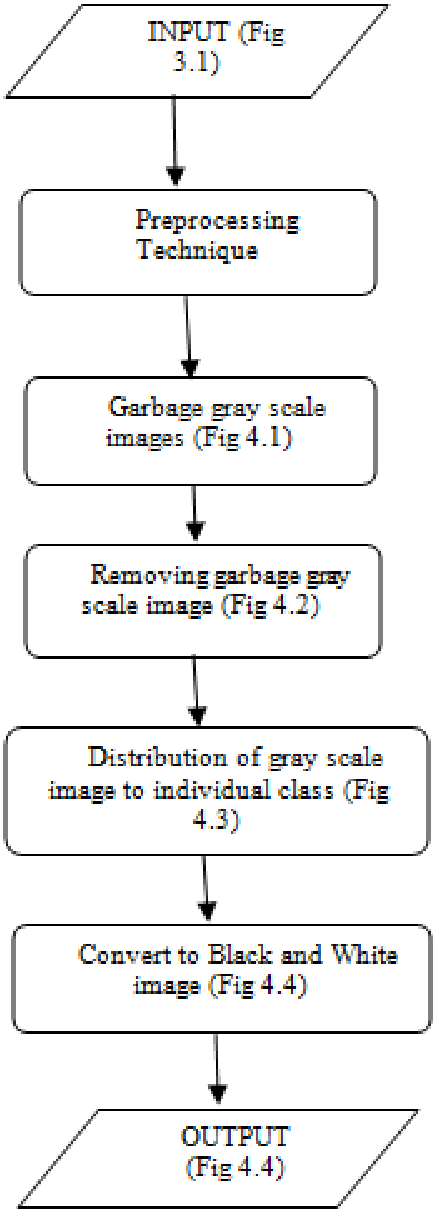

INPUT (Fig 3.1)
Preprocessing Technique
Garbage gray scale images (Fig 4.1)
Removing garbage gray scale image (Fig 4.2)
Distribution of gray scale image to individual class (Fig 4.3)
Convert to Black and White image (Fig 4.4)
OUTPUT (Fig 4.4)

Fig. 3.2 Steps of the proposed methodology.

Algorithm 1: Find the Contours and Crop

```
Read all the images from scanned directory:
    Convert the image to grayscale
    Detect Canny edge(Select the threshold)
    Find all the contours
    Make the folder for individual grayscale image
    For all contour do:
        Detect all the contours in the image
        Select all the characters(contours)
        Crop the images on detected contours area.
        Save the images into the previously created folder
    End for
End for
```

Algorithm 2: Removing Garbage Images

```
Read all the images from directory:
    Find the lowest dimension of accepted crop images
    If image dimension greater than lowest dimension
    Then save the images into the new folder
End for
```

Algorithm 3: Remove the Noise and Invert

```
Read all the images from directory:
    Resize the images(100x100) and invert it
    Set the threshold
    Fit the images into 70x70 pixel box
    Remove every row and column at the sides of the images which are black
    Resize outer box into 70x70 pixel box
    Add the 30x30 pixel black rows and columns
    Save the file into new folder
End for
```

The dataset is consist of 24582 images of 82 handwritten compound characters so they are classified into 82 classes. Initially we have splitted the dataset of 24582 data into training dataset and test dataset. The training set contains 20482 and test set complies of 4100 data. Every single image is resized as 28 X 28 pixel. All the 82 classes were labeled. The classes were named from 1 to 82 in ascending order. The class name are labeled as (class name: Label number) '1':1,'2':2. After labeling, the images were then converted into gray scale images. All images and their corresponding labels were taken in numpy array. To feed the data into machine learning algorithms, every single images were normalized. Then we converted the images into "one hot encoding" for better accuracy. In the next step, the images along with their label were fit in into the model. The proposed model is given:

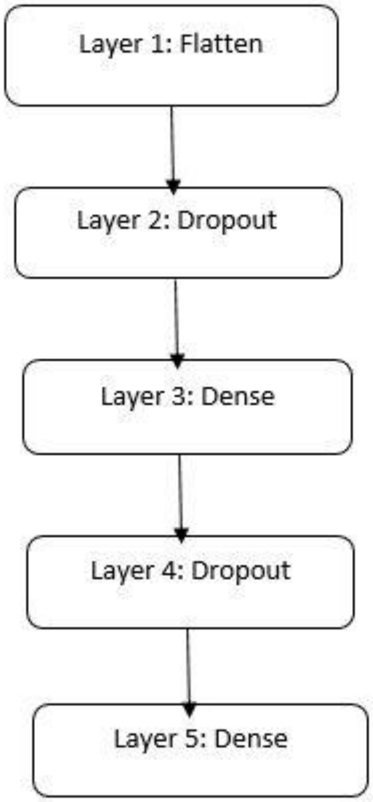


Fig. 3.3 Proposed architecture

To train the model, we have used optimizer (adam) and categorical cross-entropy as loss function and 20 times epoch and 40 times steps per epoch. After training, we evaluate the model using test data with 40 steps.

Firstly, the images of 28X28 pixel is taken into flatten layer as input. For reducing overfitting we used 25% dropout. Then we used dense or fully connected layer of 512 hidden nodes with ReLU activation followed by 25% dropout. Final output layer has 82 nodes with SoftMax activation.

3.5 Statistical Analysis

While collecting every contributor's unique handwriting, their age, gender, and educational level were also recorded. 42.11% of the total data was from female contributors and 57.89% was from male (Fig. 3.3). The maximum number of data was collected from age of 12, 13 and 14 (Fig. 3.4) [5],[8].

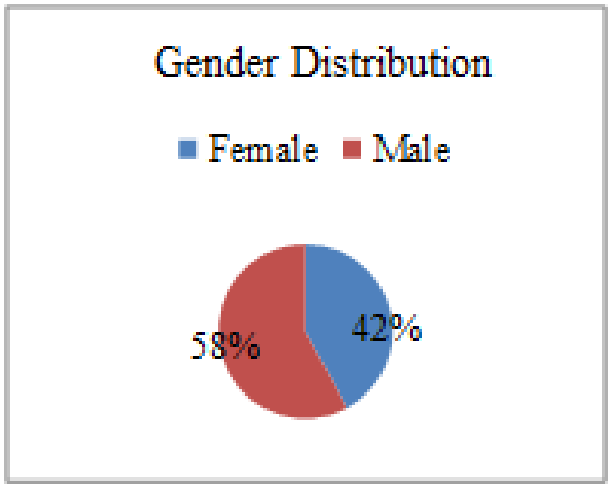


Fig. 3.4 Gender distribution of contributors.

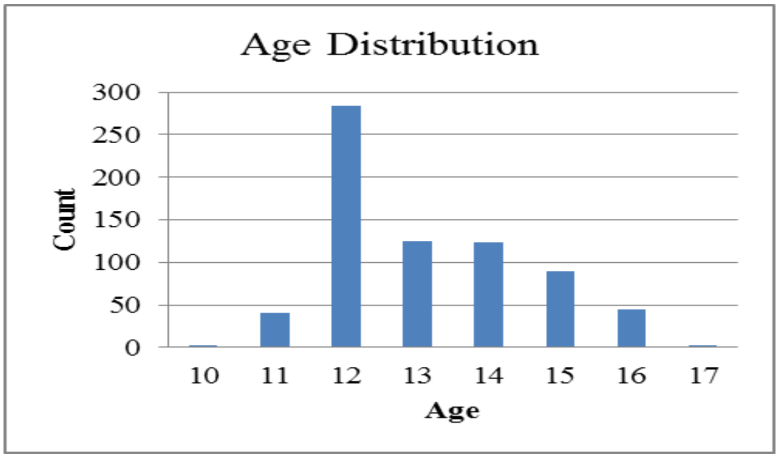


Fig. 3.5 Age distribution of contributors.

Chapter 4

Experimental Results and Discussions

4.1 Introduction

In this research, all the raw data was collected in forms from the contributors (Fig. 3.1). The selected 76 compound characters and three modifiers were provided in form so that the contributors can easily attain in this work.

4.2 Experimental Results

All the forms were scanned at 200dpi and each character of the form are extracted through OpenCV computer vision library. As the data was cropped from the forms using an algorithm, so it generated some garbage along with the necessary and usable data (Fig. 4.1). Before distributing the data, garbage data were needed to be discarded (Fig. 4.2). Then all the data were distributed to their specific classes manually (Fig. 4.3).

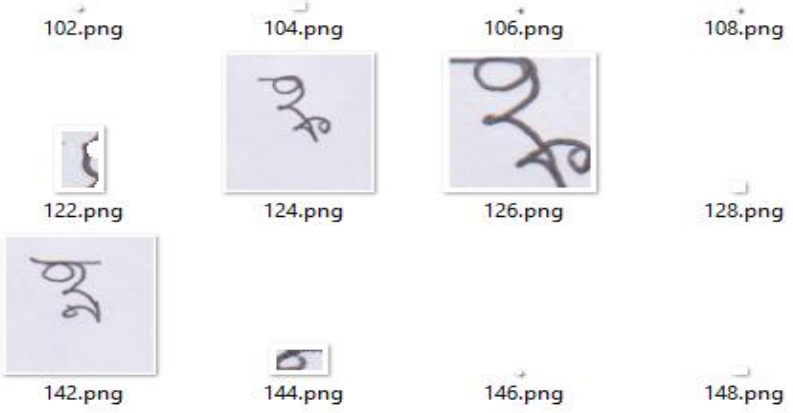


Fig. 4.1 Gray scale image with garbage.

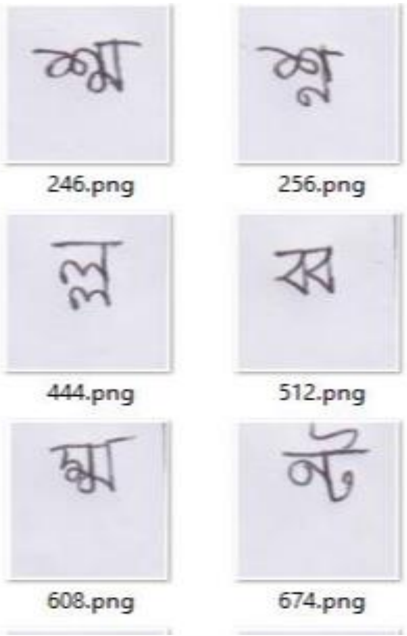


Fig. 4.2 Gray scale image.

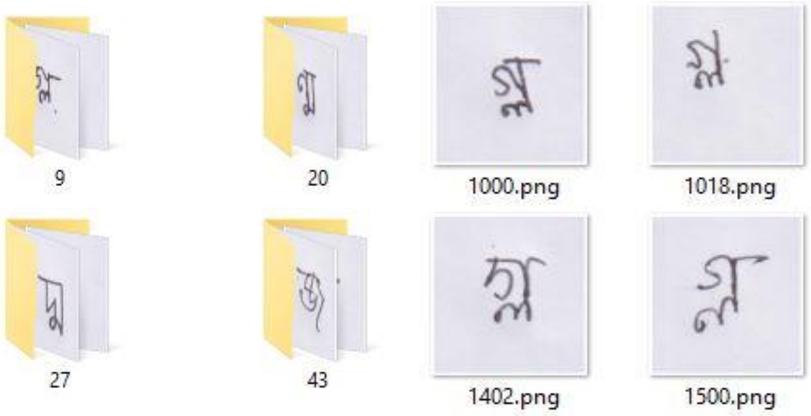


Fig. 4.3 Distribution of gray scale image to individual class.

The scanned images were consisting of white background and black writing appears. But for the machine learning task, the grayscale image was not suitable. So, the images were converted into the black background along with white characters. For the feeding the images into the model, we needed to convert the images into 100X100px (Fig. 4.4). The images were not centered efficiently. So, the images were centered using the center of mass technique.

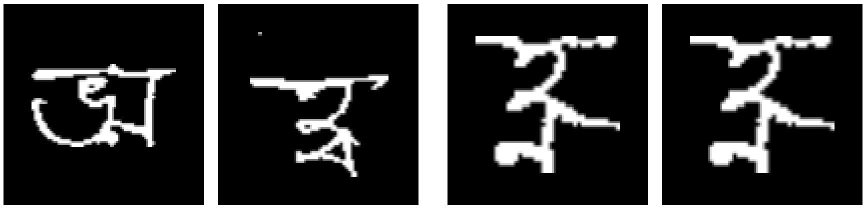


Fig. 4.4 Processed images from dataset.

After training the model we get both training and test accuracy for Bangla52 dataset. The accuracy for training dataset (20482) is 83.47% and accuracy for testing dataset (4100) is 78.73%.

4.3 Descriptive Analysis

However, there does not exist a standard relevant database, we developed a standard representative database for our Bangla handwritten compound character recognition. Requirement of such recognition technology has gained further important with easily availability in personal or portable computer device. This processed images has a great value to enrich our research data. To make all the images useable, we have to edit them, convert them, so that we can train the computer easily.

4.4 Summary

With the spread of computers in all parts of the world including developing countries need a proper documentation of their all kind of papers, that’s why it has a great importance of OCR. As like as Bangla OCR is important for us. The result of the preprocessed data is used to make the purpose easy. To implement this technology it is important to make all the data be changed according to the requirement of the algorithm. After scanning the images, the experiment is on its way to process all the data.

CHAPTER 5

Conclusion, Limitations & Future Works

5.1 Summary of the Study

From this research, we get a proper use of Bangla compound characters to computerize them. All the process is made to make the accuracy rate of Bangla OCR to be enriched. We have tried our best to emphasize on a way or method of Bangla character recognition in the simplest possible manner. There is a huge area to research on Bangla character and its recognition procedure.

5.2 Conclusions

In this paper, a large dataset of Bangla handwritten compound character with the complete approach of making a dataset for isolated characters of any language is proposed, discussed and published. The approach that we discussed above in this research to process dataset is competent to advance acceptable accuracy in machine learning model.

5.3 Recommendations

We are the nation of Bengali. So we should work continuously to make our Bangla language enriched and it is the great recommendation. We can see in every step Bangla language is lagging behind even in use of technology. We say that Bangla is a hard language, but it is not impossible. We have to try our best to make our language useable everywhere of the world. That's why Bangla is recommendable in everywhere.

5.4 Limitations and Future Works

More data could be collected to enrich our dataset. There are so many compound characters but all characters weren’t taken for our work. Data could be taken from all general people but we have selected students to make our work easy.

In near future, we would like to do further enhancement of the proposed dataset by adding more than 2000 contributors from various demographics.

References:

[1] J.G. Carbonell, R.S. Michalski, T.M. Mitchell An overview of machine learning, Machine Learning, Springer, Berlin Heidelberg (1983), pp. *3-23*

[2] S.Impedovo, More than twenty years of advancements on frontiers in handwriting recognition, Pattern Recognition. 47(2014) 916–928.

[3] Y.LeCun, Y.Bengio, G.Hinton, Deep learning,Nature 521(2015) 436–444.

[4] U.Bhattacharya, B.B.Chaudhuri, Handwritten numeral databases of Indian scriptsand multistage recognition of mixed numerals, IEEE Trans. Pattern Anal. Mach. Intell. 31.3 (2009) 444–457.

[5] Data:N.Das, A.Kallol, S.Ram, B.Subhadip, K.Mahantapas, N.Mita, A benchmark image database of isolated Bangla handwritten compound characters, Int.J. Doc. Anal. Recognition (IJDAR) 17 (4) (2014) 413–431.

[6] D.Wan,M.Zeiler, S.Zhang, S.,Y.LeCun., & R.Fergus. Regularization of neural networks using dropconnect .in:Proceedings of the 30th International Conferenceon MachineLearning (ICML-13), pp. 1058–1066.

[7] 3FingersHandwritingDevelopmentAcademy. ⟨http://3fingershandwriting.com/handwritingtips.html⟩. (accessed23.02.17).

[8] Mithun Biswas, RafiqulIslam, GautamKumarShom, Md. Shopon, Nabeel Mohammeda, SifatMomena, AnowarulAbedin, BanglaLekha-Isolated: A multi-purpose comprehensivedataset of Handwritten Bangla Isolated characters (Volume 12, June 2017, Pages 103-107), doi: https://doi.org/10.1016/j.dib.2017.03.035

[9] "Cmaterdb - CMATERdb: The pattern recognition database repository," http://www._ndbestopensource.com/product/cmaterdb, accessed: 2017-02

[10] 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, 2009.

[11] “BanglaLekha-Isolated: A Comprehensive Bangla HandwrittenCharacterDataset“https://data.mendeley.com/datasets/hf6sf8zrkc/2

[12] El-Sherif, E. & Abdleazeem, S. “A Two-Stage System for Arabic Handwritten Digit Recognition Tested on a New Large Database”.*International Conference on Artificial Intelligence and Pattern Recognition, AIPR-07.* Orlando, Florida, USA, 2007.

[13] Jawad H AlKhateeb ,“A Database for Arabic Handwritten Character Recognition” . International Conference on Communication, Management and Information Technology (ICCMIT 2015)

[14] “CASIA Online and Offline Chinese Handwriting Databases”, http://www.nlpr.ia.ac.cn/databases/handwriting/Home.html

[15] COM Offline Dataset- An Urdu Handwritten Dataset Generation”,https://iajit.org/index.php?option=com_content&task=view&id=1136

[16] “HP Labs India Indic Handwriting Datasets”, http://lipitk.sourceforge.net/hpl-datasets.htm

[17] “NIST Special Database 19”, https://www.nist.gov/srd/nist-special-database-19

[18] Fujisawa, H.: Forty years of research in character and document recognition—an industrial perspective. Pattern Recognit. 41(8),2435–2446 (2008)

[19] Cheriet, M., El Yacoubi, M., Fujisawa, H., Lopresti, D., Lorette, G.:Handwriting recognition research: twenty years of achievement and beyond. Pattern Recognit. 42(12), 3131–3135 (2009)

[20] Su, T.-H., Zhang, T.-W., Guan, D.-J., Huang, H.-J.: Off-line recognition of realistic chinese handwriting using segmentation-free strategy. Pattern Recognit. 42(1), 167–182 (2009)

[21] Srihari, S., Yang, X., Ball, G.: Offline chinese handwriting recognition: an assessment of current technology. Front. Comput. Sci. China 1(2), 137–155 (2007)

[22] Kimura, F.: OCR Technologies for machine printed and hand printed Japanese text. In: Chaudhuri, B.B. (ed.) Digital document processing. Advances in pattern recognition, pp. 49–71. Springer, London (2007)

[23] Kwon, J.-O., Sin, B., Kim, J.H.: Recognition of on-line cursive korean characters combining statistical and structural methods. PatternRecognit. 30(8), 1255–1263 (1997)

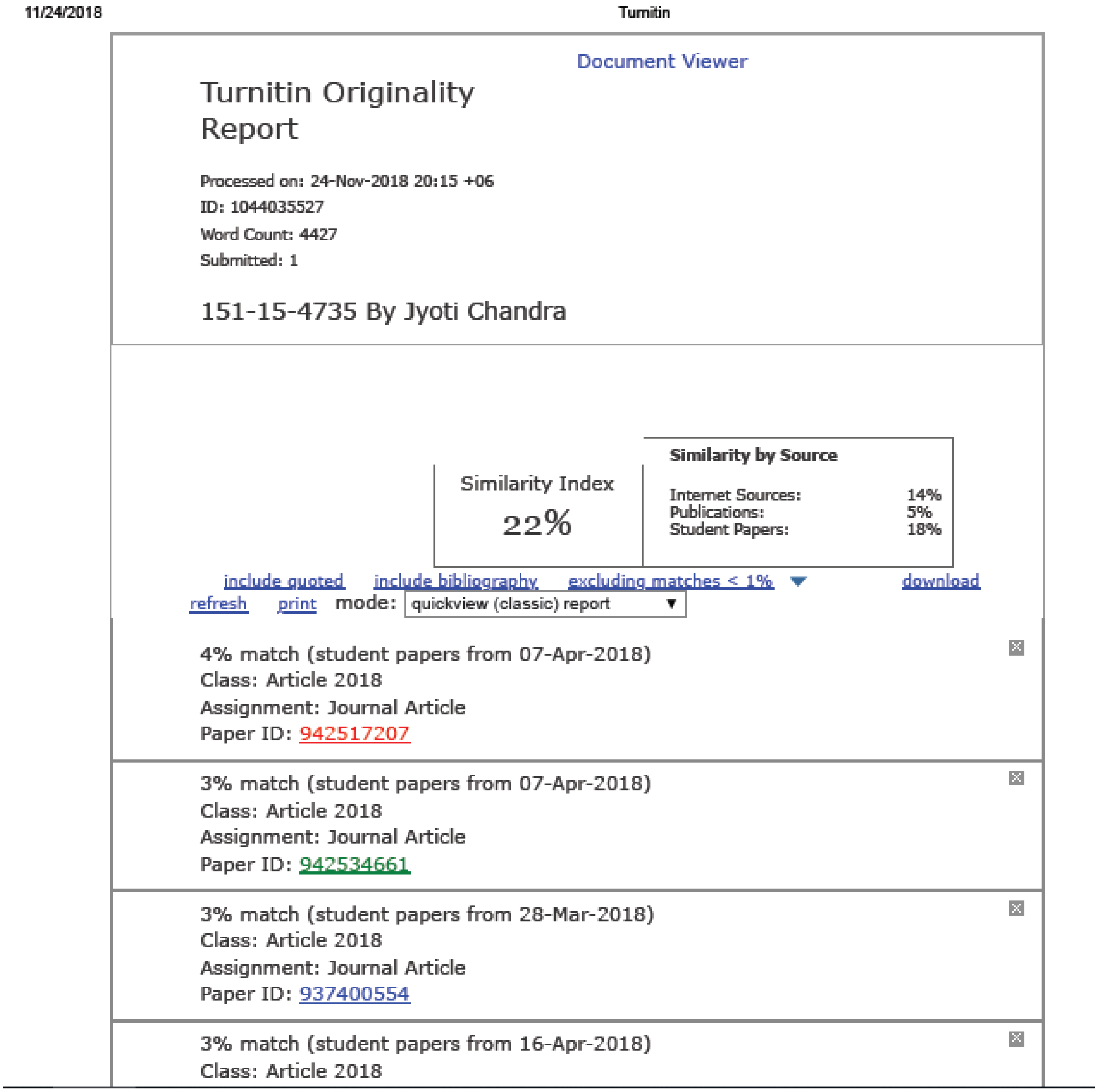
[24] Kim, H.J., Kim, P.K.: Recognition of off-line handwritten Korean characters. Pattern Recognit. 29(2), 245–254 (1996)

[25] Amin, A.: Off line Arabic character recognition: a survey. In: The fourth international conference on document analysis and recognition, pp. 596–599 (1997)

[26] Pal, U., Chaudhuri, B.B.: Indian script character recognition: a survey. Pattern Recognit. 37(9), 1887–1899 (2004)

[27] Pal, U., Jayadevan, R., Sharma, N.: Handwriting recognition in indian regional scripts: a survey of offline techniques. ACMTrans. Asian Lang. Inf. Process. 11(1), 1–35 (2012)

Plagiarism Report:



11/24/2018 Turnitin

Document Viewer

Turnitin Originality Report

Processed on: 24-Nov-2018 20:15 +06
ID: 1044035527
Word Count: 4427
Submitted: 1

151-15-4735 By Jyoti Chandra

Similarity Index	Similarity by Source	
22%	Internet Sources:	14%
	Publications:	5%
	Student Papers:	18%

4% match (student papers from 07-Apr-2018)
Class: Article 2018
Assignment: Journal Article
Paper ID: 942517207

3% match (student papers from 07-Apr-2018)
Class: Article 2018
Assignment: Journal Article
Paper ID: 942534661

3% match (student papers from 28-Mar-2018)
Class: Article 2018
Assignment: Journal Article
Paper ID: 937400554

3% match (student papers from 16-Apr-2018)
Class: Article 2018