Data in Brief 34 (2021) 106633

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

BanglaWriting: A multi-purpose offline Bangla handwriting dataset

M.F. Mridha*, Abu Quwsar Ohi, M. Ameer Ali, Mazedul Islam Emon, Muhammad Mohsin Kabir

Department of Computer Science & Engineering, Bangladesh University of Business & Technology, Dhaka, Bangladesh

ARTICLE INFO

Article history:
Received 21 October 2020
Revised 26 November 2020
Accepted 3 December 2020
Available online 9 December 2020

Keywords:
Writer identification
Word segmentation
Optical word recognition
Optical character recognition

ABSTRACT

This article presents a Bangla handwriting dataset named BanglaWriting that contains single-page handwritings of 260 individuals of different personalities and ages. Each page includes bounding-boxes that bounds each word, along with the unicode representation of the writing. This dataset contains 21,234 words and 32,787 characters in total. Moreover, this dataset includes 5,470 unique words of Bangla vocabulary. Apart from the usual words, the dataset comprises 261 comprehensible overwriting and 450 handwritten strikes and mistakes. All of the bounding-boxes and word labels are manually-generated. The dataset can be used for complex optical character/word recognition, writer identification, handwritten word segmentation, and word generation. Furthermore, this dataset is suitable for extracting age-based and gender-based variation of handwriting.

* Corresponding author.
E-mail addresses: firoz@bubt.edu.bd (M.F. Mridha), quwsarohi@bubt.edu.bd (A.Q. Ohi), dmaa730@gmail.com (M.A. Ali), emon.bubt3382@gmail.com (M.I. Emon), m97kabir2@gmail.com (M.M. Kabir).

https://doi.org/10.1016/j.dib.2020.106633

Specifications Table

Subject	Computer Vision and Pattern Recognition
Specific subject area	Optical character recognition, word segmentation, writer identification
Type of data	Image and JSON
How data were acquired	The images of the handwriting were captured using scanners and smartphone cameras. Each of the handwriting-images was cropped and annotated manually.
Data format	Raw data Converted data Annotations
Parameters for data collection	Scanner: HP Scanjet 2400 Smartphone camera: Xiaomi Redmi 6, Xiaomi Redmi 7. A single image contains the handwriting of an individual. Each individual is identified using age, gender, and unique person id. The handwritten words are segmented using bounding-boxes. Each of the bounding-boxes contains the characters that are written. Labelme [1] software is used to draw and label the bounding-boxes.
Description of data collection	The writings were conducted using regular stationery products. Writers were advised to write on a random topic. Only one page of writing was collected from each individual. The handwritings were further captured using scanners and smartphone cameras. Each captured image was cropped and annotated manually.
Data source location	Institution: Bangladesh University of Business & Technology District: Dhaka, Kishoreganj, Gopalganj, Comilla, Gazipur, Tangail, Netrakona, Mymensingh Country: Bangladesh
Data accessibility	Repository name: Mendeley Data identification number: 10.17632/r43wkvdk4w.1 Direct URL to data: https://data.mendeley.com/datasets/r43wkvdk4w/1

Value of the Data

- The dataset exploits possibilities and usage of handwritings from scanned and pictured documents. The usage of scanned and pictured forms in the recognition and identification process is often termed as an offline approach.
- The dataset is suitable for machine learning [2] models, deep learning [3] models, producing embedding vectors [4] of handwriting, etc.
- The dataset exploits all possible potentials of Bangla handwriting [5]. The dataset contains bounding-box annotations for each handwritten word, unicode representation for each written word, and writer information for each document. Therefore, the dataset is suitable for word segmentation, optical character recognition, writer identification, writer verification, and handwriting generation.
- The dataset contains raw images (without any pre-processing) of each document. The dataset also contains supplementary pre-processing scripts to suspend excess lighting and noises.
- The dataset can be used to explore writing patterns related to age and gender.

1. Data Description

BanglaWriting, the dataset presented in this paper, aims to provide a preferable handwriting dataset that is enriched from every dimension. The dataset can be used in diverse machine learning and deep learning based applications. It can be implemented in handwriting biometric tasks, including identification, verification, and age/gender estimation. Further, the dataset has possibilities for specific computer vision tasks such as optical character recognition and handwriting segmentation. Moreover, the dataset has the capability of fueling generative handwriting models. Fig. 1 illustrates the possible domains on which the dataset can contribute.

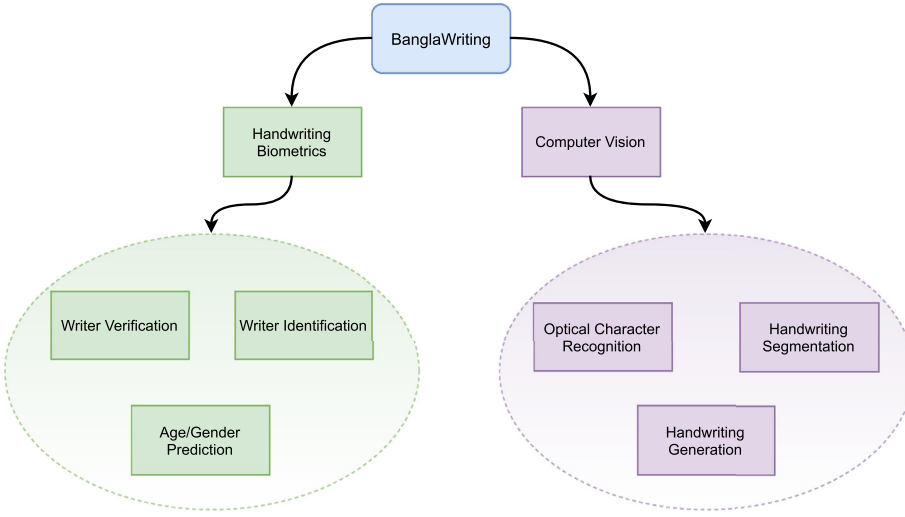


Fig. 1. The BanglaWriting dataset can be used for handwriting biometrics and computer vision-specific tasks. The dataset has possibilities in various fields, including identifying writers, to generating handwritings from unicode.

Table 1
The table illustrates a quantitative comparison of the BanglaWriting dataset with some famous datasets in different languages. The BanglaWriting dataset targets almost all possible domains of interest in offline handwriting processing. In general, most datasets neglect various classes (overwriting, random strikes) of handwriting. Hence, we exclude the number of classes in comparison.

Dataset	Language	Writers	Total Documents	Word Count	Word-level Bounding-box
RIMES [8]	French	1300	12723	300000	Yes
KHATT [7]	Arabic	1000	2000	165890	No
IAM [5]	English	400	1066	82227	Yes
BanglaWriting	**Bangla**	**260**	**260**	**21234**	**Yes**
Firemaker [9]	Dutch	252	1008	-	No
AHDB [10]	Arabic	105	-	10000	Yes

This dataset's construction and usage are different from usual Bangla datasets [6]. The currently available datasets for Bangla writing only include isolated character writings. Whereas, the BanglaWriting dataset contains word-based writing with bounding-boxes. The dataset is implemented based on well-known offline handwriting, and writer recognition datasets [5]. Table 1 presents a comparison BanglaWriting dataset with some of the popular datasets of diverse languages. Most of the bigger datasets (such as KHATT [7], IAM [5]) include some automated and pre-estimated parameters to label the data. In comparison, the annotations and labels of the BanglaWriting dataset are manually determined. Hence from the overall evaluation, it can be concluded that the BanglaWriting dataset attains a marginal amount of quality data.

The BanglaWriting dataset contains single-page handwritings of 260 individuals from eight different districts (illustrated in Table 4). It consists of 5,470 unique words and 124 unique characters. Moreover, the overall dataset comprises 21,234 words and 32,787 characters in total. The dataset contains Bangla characters, numerics, diacritics, and conjuncts. Furthermore, it has punctuation marks and English alphabets mixed with Bangla writing. Table 2 illustrates the Bangla characters that exist in the dataset. For better understanding, Fig. 2 explicates the underlying construction of a Bangla word. Fig. 3 illustrates a sample of the BanglaWriting dataset, bounding-box, and labels.

Table 2
The BanglaWriting dataset contains all characters of Bangla vocabulary. The table illustrates the Bangla characters that also exist in the dataset.

Character Type	Characters
Vowel	অ, আ, ই, ঈ, উ, ঋ, এ, ঐ, ও, ঔ
Consonant	ক, খ, গ, ঘ, ঙ, চ, ছ, জ, ঝ, ঞ, ট, ঠ, ড, ঢ, ণ, ত, থ, দ, ধ, ন, প, ফ, ব, ভ, ম, য, র, ল, শ, ষ, স, হ, ৎ, ড়, ঢ়, য়
Diacritic	া, ি, ী, ু, ূ, ৃ, ে, ৈ, ো, ৌ, ্
Numeral	০, ১, ২, ৩, ৪, ৫, ৬, ৭, ৮, ৯

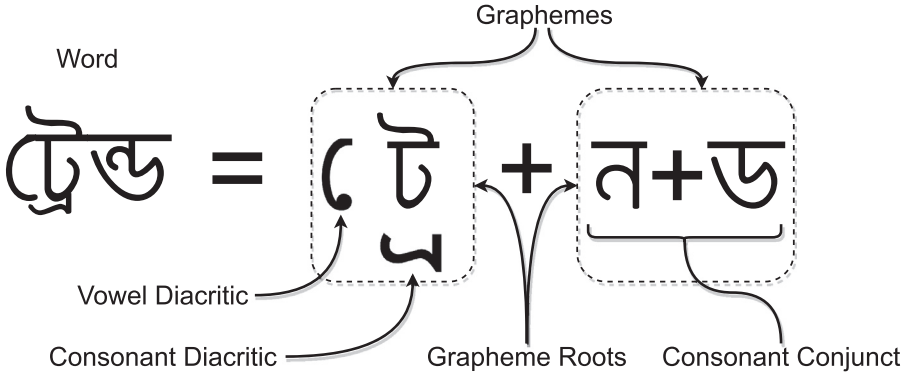
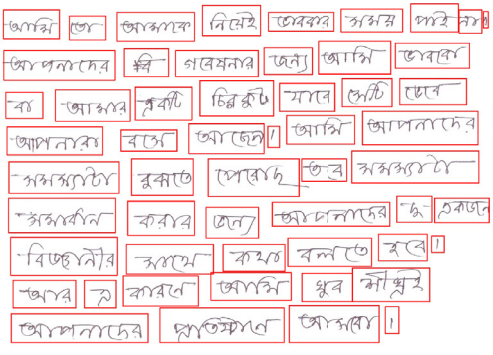


Fig. 2. Graphemes are the smallest unit of meaningful writing. A grapheme always contains a grapheme root. In the Bangla writing system, a grapheme may have one vowel and one consonant diacritic. Occasionally, a grapheme may include consonant conjuncts as it's grapheme root.



আমি তো আমাকে নিয়েই ভাববার সময় পাই না।
আপনাদের * গবেষনার জন্য আমি ভাববো
বা আমার একটি চিরকুট যাবে সেটি ভেবে
আপনারা বসে আছেন। আমি আপনাদের
সমস্যাটা বুঝতে পেরেছি তবে সমস্যাটা
সমাধান করার জন্য আপনাদের দু একজন
বিজ্ঞানীর সাথে কথা বলতে হবে।
আর এ কারণে আমি খুব শীঘ্রই
আপনাদের প্রতিষ্ঠানে আসবো।

Fig. 3. The left image illustrates a handwriting image with word-level bounding-boxes. The labels/words for each bounding-box is presented on the right. The excluded word (second row, second word) is marked using an asterisk (*).

The dataset is presented in two different versions, (i) raw and (ii) converted. The raw file contains raw images that were manually cropped, and no image-processing techniques were applied. Hence, the raw dataset includes a diversity of color shifts, shadowing effects in images. On the contrary, the converted file contains a furnished version of the raw images (discussed in Section 2.5). Fig. 6 illustrates the difference between the raw and converted dataset images. Further, Fig. 4 shows the directory structure for both dataset versions.

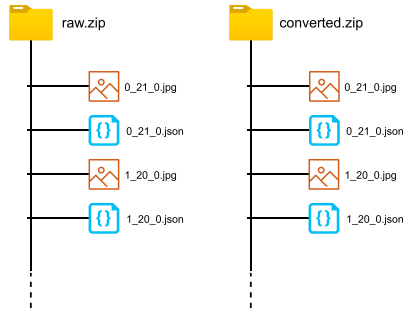


Fig. 4. The figure illustrates the directory structure of the BanglaWriting data files. The 'raw.zip' contains raw images that were only labeled. The 'converted.zip' contains labels, and the images are manually processed using the additional script [11]. For every image file, there exists a *JSON* file with the same naming scheme. The *JSON* file contains the bounding-boxes and labels.

Table 3
The table describes the quantitative distribution of each label along with the labeling schemes.

Class	Count	Label Scheme
Clear writing	21234	Contains word in unicode
Overwriting	261	Contains word in unicode with asterisk '*'
Strike and mistakes	450	Contains asterisk '*'

Each of the pictures in both datasets comprises the writing of a single individual. Individual images are named based on the following convention,

$$personIdentifier\_age\_gender.jpg$$

Where,

$personIdentifier$ = Unique id assigned to an individual.
age = Age of the individual.
$gender$ = Gender of the individual. 0 for females and 1 for males.

For every image data, a *JSON* file is also included with the same naming convention. The *JSON* file contains the word-level bounding-box information and labels for each bounding-box. The *JSON* format is illustrated in Fig. 9 and it is further elaborated in Section 2.4.

The labels for each word-level bounding-box represents the words written in unicode format. There are three possible classes/label-formats maintained, which are presented below.

1. **Clear writing**: By clear writing, we refer if the bounding-box contains written word that the writer intended to write and are understandable. In this case, we label the bounding-box with the unicode value of the written word.
2. **Overwriting**: By overwriting, we refer if the bounding-box contains the written word, but some of the characters have been stroked out. Writers often strike-out some character to refer to exclude that character. In such a case, we label the comprehensible characters with proper unicodes, and we omit the stroked out characters in the label. In such a case, we add an asterisk ('*') with the Unicode label to mark the issue.
3. **Strikes and mistakes**: The dataset contains some random strikes (such as word underlines, rules), and fully stroked out words. We do not include any unicode in such cases, and we only label them using an asterisk ('*').

Fig. 5 further illustrates some examples of the labels mentioned above. Moreover, Table 3 represents the quantitative distribution of each class in the dataset.

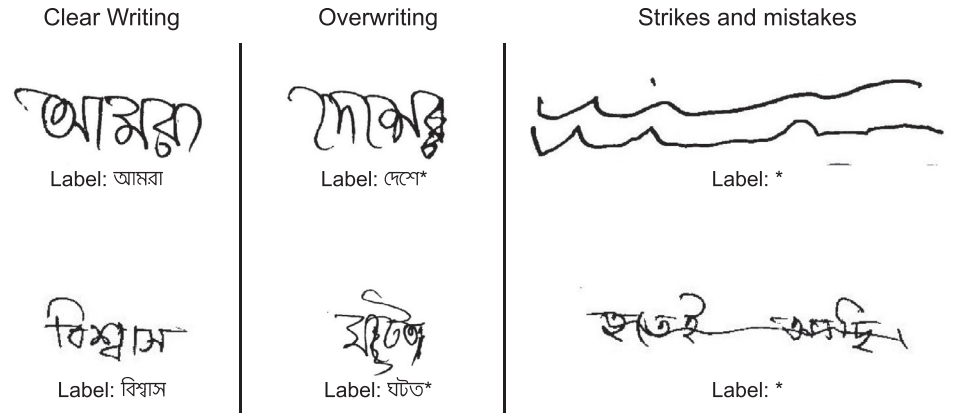


Fig. 5. The figure depicts some examples of the words and labels generated for each class. The left, middle, and right columns explicate clear writing, overwriting, and strikes/mistakes, respectively.

Table 4
The table describes the quantitative distribution of the geographical location of the writers.

District	Total Documents
Dhaka	48
Gopalganj	26
Comilla	14
Gazipur	21
Tangail	36
Netrakona	25
Kishoreganj	46
Mymensingh	44

The dataset also includes a supplementary script [11] used to produce the furnished images of the 'converted' version of the data. The script is used to reduce the noises and light variations of the 'raw' data images.

2. Experimental Design, Materials and Methods

2.1. Data collection

The dataset was collected from the students of Bangladesh University of Business and Technology. Furthermore, to generate a better age distribution of the dataset, the students' household members were also included. Fig. 7 illustrates the age and gender distribution of the population. However, the writers were selected based on the primary clinical constraints, (a) The minimum age of the writers can be 8, (b) The writers should be physically fit to write.

The writers written on A4-sized papers, and regular ball-point and gel pens were used for writing. Each individual was suggested to write on any topic. Therefore, each document contains a diverse number of words. Fig. 8 represents the word distribution per document. Moreover, allowing writers to write on random topics also resulted in making mistakes and overwriting that are also labeled.

The writers are from eight different districts of Bangladesh. We define a writer belonging to a particular district if he/she stayed in the district for more than ten years. Table 4 illustrates a quantitative distribution of the geographical location of the writers.

2.2. Data extraction

The handwritten pages were further imaged using a scanner and smartphone cameras. The dataset contains a total of 52 scanned images and 208 images captured using smartphone cameras. The scanned images do not contain any noisy conditions. On the contrary, the images captured using smartphone cameras have noises due to environmental factors, such as various lighting effects, glazes of flashlight, and shadow effects.

2.3. Data preprocessing

Each image data were cropped and strengthened manually. The images were named using the formula, *personIdentifier_age_gender*. No augmentation was applied to increase the dataset's size to ensure the dataset's authenticity and quality.

2.4. Data labeling

The dataset was manually annotated using *labelme* [1] software. Fig. 3 illustrates the word-based bounding-boxes and the unicode-text labels for each bounding-box. The figure also demonstrates the annotation policy adapted for overwriting and cropped words/characters. Table 3 illustrates the labeling policy adopted for three different labels/classes of the word-based bounding-boxes.

The bounding-box and label information for each image was separately saved on individual JSON files, following the same naming convention of the handwritten images. Fig. 9 illustrates the standard JSON-file parameters that were generated for each image. The *"shape"* property contains an array of *"label"* and *"points"* parameter pairs. The *"label"* parameter contains the written word (in unicode-8) in the bounding-box. Whereas, the *"points"* parameter contains an array of starting and ending pixel-coordinates of the bounding-box. The *"imagePath"*, *"imageHeight"*, and *"imageWidth"* contains some additional information such as, the filename of the corresponding image, the height and width of the image, respectively.

2.5. Supplementary script

As the dataset contains raw images taken using scanners and smartphones, a difference of lightning and background noise is noticed (illustrated in Fig. 6). Hence, the dataset includes a supplementary *Python* [12] and *OpenCV* [13] based script [11] that eliminates lightning issues and reduces the background noises. The script further furnishes the images and generates images suitable for machine learning and deep learning strategies. The furnished images are provided in the 'converted.zip' file, whereas the 'raw.zip' contains the raw images where no image-processing techniques were applied.

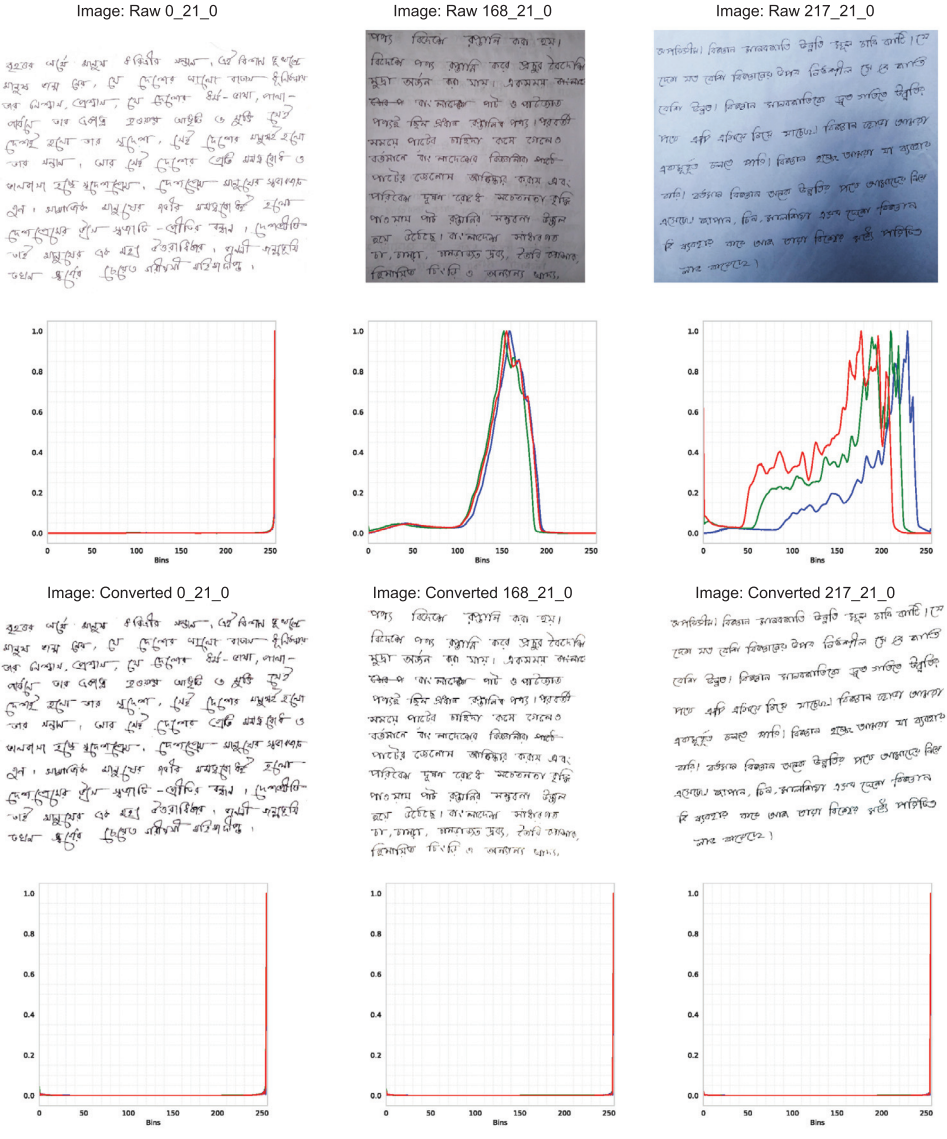


Fig. 6. The illustration points out the image data variation in the 'raw' and 'converted' versions of the BanglaWriting dataset. The upper row illustrates the raw version's image data, where the first image is taken using a scanner, and the rest are captured using a smartphone camera. The second row illustrates color histograms w.r.t. the images. The third row depicts the same pictures from the converted version (processed using the supplementary script [11]). The fourth row illustrates the color histogram w.r.t. the images in the third row. By comparing the color histograms, it can be concluded that the 'raw' version's images contain color shifts and light issues. In contrast, the converted images exclude those challenges.

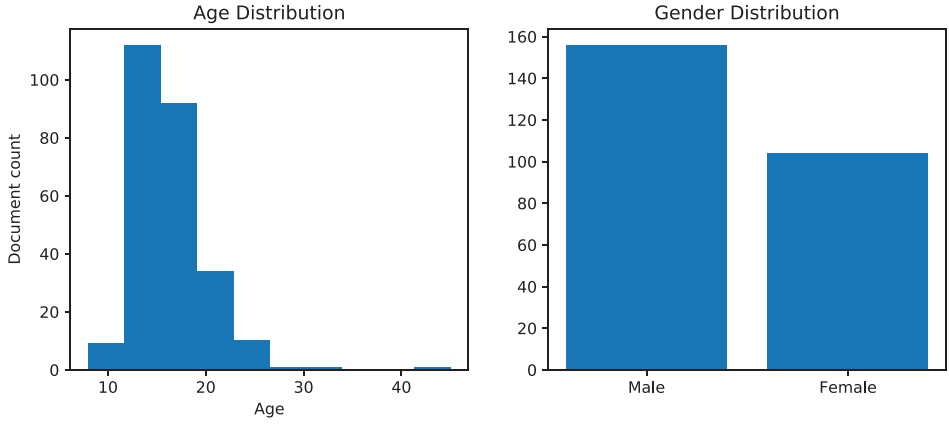


Fig. 7. The left graph exhibits age distribution, and the right graph demonstrates the gender distribution of the dataset.

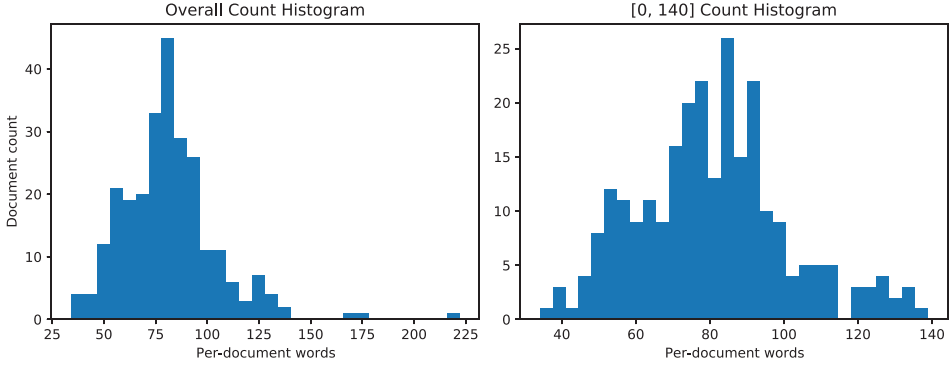


Fig. 8. The left graph illustrates the word per document distribution for each paper. The right graph shows the same scenario without outliers. The word-count histogram simulates normal distribution.

```
{
        "shapes": [
        {"label": "wordLabel",
         "points": [[xmin, ymin], [xmax, ymax]]
        },
        {"label": "wordLabel",
         "points": [[xmin, ymin], [xmax, ymax]]
        },
        ....
        ],
        "imagePath": "uniquePersonIdentifier_age_gender.jpg",
        "imageHeight": Xpx,
        "imageWidth": Ypx
}
```

Fig. 9. The figure illustrates a JSON structure that interprets the bounding-boxes and labels information for each handwriting image data.

Ethics Statement

All the handwritings were obtained with the consent of the individuals who had participated in the writing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

Acknowledgments

The authors would like to thank the Advanced Machine Learning (AML) lab and the Bangladesh University of Business and Technology (BUBT) for their resource sharing and precious suggestions.

References

[1] K. Wada, labelme: Image Polygonal Annotation with Python, 2016, (https://github.com/wkentaro/labelme).
[2] D. Michie, D.J. Spiegelhalter, C. Taylor, et al., Machine learning, Neural Stat. Classif. 13 (1994) (1994) 1–298.
[3] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, nature 521 (7553) (2015) 436–444.
[4] A.Q. Ohi, M. Mridha, F.B. Safir, M.A. Hamid, M.M. Monowar, Autoembedder: a semi-supervised DNN embedding system for clustering, Knowl.-Based Syst. 204 (2020) 106190.
[5] U.-V. Marti, H. Bunke, The iam-database: an english sentence database for offline handwriting recognition, Int. J. Document Anal. Recognit. 5 (1) (2002) 39–46.
[6] M. Biswas, R. Islam, G.K. Shom, M. Shopon, N. Mohammed, S. Momen, A. Abedin, Banglalekha-isolated: a multi-purpose comprehensive dataset of handwritten Bangla isolated characters, Data in Brief 12 (2017) 103–107.
[7] S.A. Mahmoud, I. Ahmad, W.G. Al-Khatib, M. Alshayeb, M.T. Parvez, V. Märgner, G.A. Fink, Khatt: an open arabic offline handwritten text database, Pattern Recognit. 47 (3) (2014) 1096–1112.
[8] E. Grosicki, M. Carre, J.-M. Brodin, E. Geoffrois, Rimes evaluation campaign for handwritten mail, Processing (2008).
[9] L. Schomaker, L. Vuurpijl, Forensic Writer Identification: A Benchmark Data Set and a Comparison of Two Systems, NICI (NIjmegen Institute of Cognitive Information), Katholieke Universiteit, 2000.
[10] S. Al-Ma'adeed, D. Elliman, C.A. Higgins, A data base for arabic handwritten text recognition research, in: Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition, IEEE, 2002, pp. 485–489.
[11] A.Q. Ohi, BanglaWriting: a multi-purpose offline Bangla handwriting dataset (script), 2020, (https://github.com/QuwsarOhi/BanglaWriting).
[12] G. Rossum, Python reference, Manual (1995).
[13] G. Bradski, The OpenCV Library, Dr. Dobb's Journal of Software Tools, 2000.