

Construction of a Tibetan Handwriting Khyug-yig Dataset

Dorje Tashi^{1**}, Tianying Sheng^{3**}, Bingtian Chen^{2**}, Renzeng Duoje¹, Rinchen Dongrub¹, Yongbin Yu^{2†},
Xiangxiang Wang^{2†}, Nyima Tashi^{1†}

¹School of Information Science and Technology, Tibet University, Lhasa 850000, China

²School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China

³School of Computer Science, University of Sydney, Sydney 2006, Australia

Keywords: Dataset; Handwriting; Tibetan; Khyug-yig Calligraphy; Text Recognition

Citation: Tashi D., Sheng T.Y., Chen B.T., et al.: Construction of a Tibetan Handwriting Khyug-yig Dataset. Data Intelligence 6(4), 870-887 (2024). doi: <https://doi.org/10.3724/2096-7004.di.2024.0048>

Submitted: April 4, 2024; Revised: July 10, 2024; Accepted: August 20, 2024

ABSTRACT

The scarcity of Tibetan handwriting datasets has hindered the applications of prevailing artificial intelligence models in the Tibetan language. As Khyug-yig is the most common writing style found in the daily lives of Tibetan people, this study proposes a methodology to construct a Tibetan handwritten Khyug-yig dataset to support further research in Tibetan fields. This approach starts with filtering the textual content of writings from multiple sources, encompassing news, medicine, and Buddhism, establishing a corpus of frequent Tibetan words. These words were organized into forms and assigned to 63 Tibetan writers across diverse institutions, including Changdu City's Sixth Senior High School, Tibet University, and a local calligraphy association. The collected handwriting forms were then processed through scanning, cropping, image preprocessing, grouping, and labeling. As a result, a Tibetan handwriting dataset with 9,874 unique-word images written in Khyug-yig style was constructed, overcoming the limitation of existing Tibetan handwriting datasets, while achieving calligraphic diversity and precise labeling.

[†] Corresponding authors: Yongbin Yu (E-mail: ybyu@uestc.edu.cn; ORCID: 0000-0001-6022-7504), Xiangxiang Wang (E-mail: xxwang@uestc.edu.cn; ORCID: 0000-0001-9341-1068), Nyima Tashi (E-mail: nmzx@utibet.edu.cn; ORCID: 0000-0001-9288-6600).

^{**} These authors contributed equally to this work.

1. INTRODUCTION

A high-quality, diverse dataset is essential for enhancing model generalizability and robustness in real-world applications. In the era of deep learning and large models, the scarcity of Tibetan handwriting data poses a great challenge across various scientific domains, such as Tibetan text recognition. To mitigate the impact of insufficient data, researchers have explored augmentation techniques, the creation of synthetic datasets, and the development of methods tailored for low-resource Tibetan data. For instance, Bao et al. (2023) [1] proposed an enhanced attentive Siamese Long Short-Term Memory (LSTM) network for Tibetan-Chinese plagiarism detection, which includes translation-based data augmentation to expand the bilingual training dataset. An (2023) [2] introduced a prompt learning-based method for low-resource Tibetan text classification, leveraging pre-trained language models to learn text representation and generation capabilities on a large-scale unsupervised Tibetan corpus. However, the paucity and limited diversity of available Tibetan data undeniably impede the development of more accurate and reliable models. As the demand for accurately labeled handwritten images in domains like pattern and optical character recognition continues to increase, collaborative efforts to construct high-quality datasets remain crucial for fundamentally addressing data scarcity and advancing research on Tibetan subjects.

1.1 Backgrounds

In Tibetan tradition, Uchen and Umê are two fundamental types of Tibetan script [3]. Khyug-yig, a subcategory of Umê, is the most prevalent, intricate, and abstract handwriting style of Tibetan calligraphy. Unlike other calligraphic styles, Khyug-yig is extremely cursive, with consonants and syllables joined together, allowing for quick handwriting. This makes it the most commonly used style for notes and personal letters.

Although Khyug-yig is ubiquitous among the Tibetan population and frequently appears in Tibetan medicine and ancient Buddhist texts, research in Tibetan calligraphy, especially in the Khyug-yig style, has been hindered due to the limited availability of well-labeled and public handwriting datasets. Thus, with the increasing need for digitalization to support Tibetan cultural preservation, medical applications, and commercial purposes, constructing Khyug-yig handwritten datasets could significantly benefit Tibetan-focused research areas; for example, text recognition of Tibetan medical prescriptions and historical archives, as well as handwriting imitation to augment sample quantity and diversity.

Existing methodologies for building Tibetan handwritten datasets primarily involve manual labeling of scanned handwriting or machine synthesis. Manual labeling processes real-world datasets through detailed human annotation of images, while machine synthesis creates handwritten images by converting input text using deep neural networks. Yang et al. (2023) [4] utilized a human-labeled dataset for text recognition in ancient Tibetan books from Dunhuang. Dhondup et al. [5] synthesized a specific Tibetan dataset with wooden pattern backgrounds and printed Tibetan text for recognizing wooden books. Similarly, Tong et al. [6] designed a Umê-style dataset for scene character recognition.

However, these approaches can either consume high labor costs or result in artificial-looking handwriting images. Additionally, following these traditional process, the constructed datasets face challenges in limited data diversity. To be specific, insufficient font features, lack of specific textual context, and restricted calligraphies, particularly for historical scriptures and wooden books, are often caused by duplicate ancient writer groups. The repetition of writers and skewness of calligraphic style can adversely affect the advancement of Tibetan models.

1.2 Previous Work

1.2.1 Image Datasets

Some commonly used image datasets have been developed to support models in computer vision, such as image generation, data augmentation, and optical character recognition (OCR). These datasets' construction methodologies are both intriguing and valuable. We have listed them for reference:

ImageNet [7]. The creation of ImageNet (Deng et al., 2009) involved two primary processes: collection and cleaning. Initially, images were web-scraped from search engines using multiple languages and queries, combined with word phrases from WordNet [8]. This approach ensured the diversity and hierarchy of the collected candidates. Subsequently, to refine the candidate pool, the Amazon Mechanical Turk (AMT) platform was utilized to assign paid annotation tasks to multiple online users. Additionally, a dynamic algorithm was developed to evaluate the confidence level of each image-label pair's accuracy. This process resulted in a powerful and widely-recognized image dataset.

ESP Database [9]. The ESP database was generated through an online interactive game where players selected image contents from a relevant label pool. This method efficiently produced numerous labeled images. However, the ESP database suffered from issues like the disambiguation of synonyms and a tendency towards basic semantics, such as 'dog', rather than more specific categories, such as 'Siberian Husky', as discussed by Rosch and Lloyd (1978) [10].

1.2.2 OCR and Handwritten Datasets

In addition to these multicategorical image datasets, the construction methodologies of several trending OCR and handwriting datasets have also garnered attention.

MNIST [11] and EMNIST Datasets [12]. The MNIST dataset (Yann LeCun and Corinna Cortes, 1994), a well-known handwritten digit dataset, was derived from the NIST dataset [13], which was sizeable but skewed with a biased distribution of writer groups across training and testing sets. Techniques such as size-normalization and centralization of position were applied to each scanned image to form the MNIST dataset, leading to high-quality, preprocessed data in the field of pattern recognition. Furthermore, the EMNIST extension by Cohen et al. (2017) introduced a wide variety of digits and letters.

IAM Dataset [14]. The IAM dataset is an English sentence dataset created by approximately 400 writers using the Lancaster-Oslo/Bergen (LOB) corpus [15]. In the data collection process, corpora were divided

into forms that organized handwritten textual content with standard spacing instructions. Each author filled these forms in their natural writing style, aiming for data fidelity and diversity. The completed forms were then scanned into images, featuring both printed labels and handwritten text, facilitating easy labeling. This approach is adaptable for constructing our Tibetan Handwriting Khyug-yig Dataset.

CASIA [16] and HIT-OR3C Databases [17]. Chinese character-level databases, such as CASIA Online Handwriting Database 1 and HIT-OR3C Chinese Character, followed similar procedures to the IAM database. CASIA comprises 3,866 characters and 171 symbols, produced by 420 writers using the Anoto pen. This diversity in characters and natural writing styles, along with different positioning within designed forms, allowed the database to overcome common issues like overly tidy handwriting or limited scale. Conversely, the HIT-OR3C Chinese handwritten corpus used a handwriting pad to record various calligraphic strokes, yielding strong pattern recognition results. It incorporates the full GB2312-80 character set of 6,763 categories, recorded by hundreds of individuals.

1.2.3 Tibetan Handwriting Datasets

Despite the maturity of English and Chinese datasets, existing labeled Tibetan handwriting datasets are limited, often produced by manual annotations of handwriting pictures or machine synthesis. In addition to these, there are several Tibetan handwriting datasets that are constructed directly by the authors writing specified content:

TibetanMNIST Dataset [18]. MNIST of Tibetan Handwriting was inspired by the renowned handwriting numeral dataset MNIST [11]. The construction of TibetanMNIST was supported by groups of Tibetan research students. After a month of writing and quality inspection, they collected 17,768 high-resolution Tibetan handwritten numeral images. Notably, it is the first Tibetan handwriting image dataset worldwide, which has been fully available to the public, significantly contributing to Tibetan handwriting recognition and other research fields.

Tibetan Handwritten Consonants Dataset [19]. Tibetan Handwritten Consonants Dataset contains a number of handwriting image samples of 30 Tibetan consonants. The construction involved 150 Tibetan students from Minzu University of China. Standard writing assignments were allocated to each participant, with provided grid forms to ensure the data quality, uniform font size, and standard writing style. The quality was controlled by manual filtering of invalid writing samples, following preprocessing techniques including color extraction, median filtering, edge detection, consonant extraction, and size normalization. As a result, 77,636 high-quality Tibetan handwritten consonant samples with approximately 2,000 images for each consonant were preserved.

1.3 Motivation and Contributions

Although some Tibetan handwriting datasets, like TibetanMNIST Dataset [18] and Tibetan Handwritten Consonants Dataset [19], have emerged with high resolution and quality, they contain only numeral digits or consonant characters, lacking content variability and domain relevance. Furthermore, issues persist in

handwriting datasets sourced from historical scriptures and wooden books, including insufficient font features and style similarity due to identical writer groups. Limited data diversity, excessive manual labor costs, and the unreliability of human annotation have restricted the emergence of Tibetan handwriting datasets and further obstructed the development of sophisticated and applicable algorithms such as Tibetan handwriting OCR. Hence, expanding Tibetan handwritten data with diverse textual content and ground-truth labels could significantly contribute to research for recognizing and generating various Tibetan scripts. Moreover, the lack of an existing public handwriting dataset for Khyug-yig, the most predominant writing form in practice, highlights the necessity of constructing such a dataset for this calligraphic style.

Inspired by these challenges, we propose a methodology to construct a Tibetan handwritten Khyug-yig dataset, which incorporates conventional techniques used in developing English and Chinese datasets. This approach specifies textual content collected from multiple sources, mainly through web crawling, encompassing news, medicine, and Buddhism, to establish a corpus of frequent Tibetan words. These words were organized into forms and assigned to 63 Tibetan writers across diverse institutions. The collected handwriting forms were then processed through scanning, cropping, image preprocessing, grouping, and labeling, which helped improve clarity and highlight the major features of the handwritten images. Ultimately, we successfully established a Tibetan handwriting dataset with 9,874 unique high-resolution images of words written in the Khyug-yig style. The dataset overcomes the limitations of existing Tibetan handwriting datasets by providing accurate ground-truth labels and addressing the lack of diversity in both textual context and calligraphic styles. Figure 1 showcases some examples from our Tibetan Handwriting Khyug-yig Dataset.



Figure 1. Example of our constructed Tibetan Handwriting Khyug-yig Dataset.

2. METHODOLOGY

2.1 Overview

The construction of the Tibetan script handwritten dataset aimed to train and evaluate recognition models. Currently, research in this area primarily utilizes two methods to build datasets: generating data with pseudo-labels and manually annotating existing text images, like scanned versions of Tibetan ancient documents. However, datasets constructed by these methods suffer from incomplete character features, a singular font style, high redundancy, and low reliability. Therefore, this paper, drawing on the construction methodologies of English and Chinese handwritten datasets, proposes a Tibetan dataset construction methodology. It begins by first determining the content, followed by selecting the subjects for collection. This approach involves studying the degree of semantics contained in various granularities of Tibetan text and the characteristics of handwriting images. It adds functionality and universality to the dataset, making it more relevant to the everyday lives of Tibetan people. Figure 2 illustrates the detailed steps of the construction process for the Tibetan handwritten Khyug-yig style dataset.

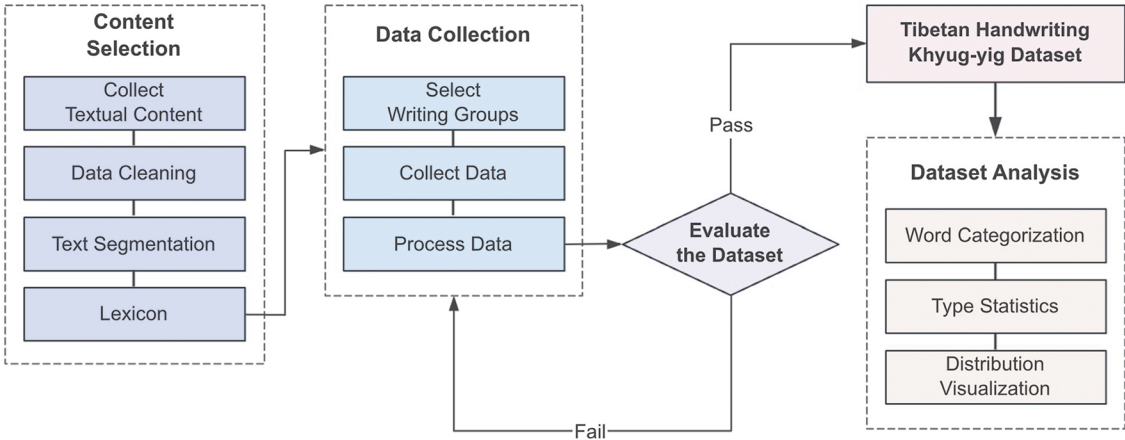


Figure 2. Flow chart for constructing Tibetan Handwriting Khyug-yig Dataset.

The scheme for establishing the Tibetan script handwritten Khyug-yig style dataset comprises three main parts: selecting candidate text content, data collection, and dataset analysis. The content selection phase primarily consists of four steps: collecting various types of Tibetan text data; performing noise reduction preprocessing, which includes converting non-Tibetan characters in the collected Tibetan text data to 'N'; conducting text segmentation; and finally, constructing a lexicon according to word frequency. The data collection phase encompasses three main tasks: the selection of capable candidates who can write in Khyug-yig style font; completing writing tasks on specified forms; and subsequent data processing. Lastly, the dataset analysis part involves calculating the distributions of each Tibetan word type to verify the effectiveness of the dataset construction.

2.2 Selecting Candidate Text Content

Selecting candidate text content was focused on ensuring the dataset's diversity and usability, particularly for the daily use of Tibetan people. This involved web scrapes and manual records. Web crawling extracted a total of 2,058.24MB of Tibetan text data from multiple websites, including the Tibet News Network, China Tibet Online, the China Tibetan Network, and the Qiongmai Literature Network. On the other hand, high-quality data amounting to 33MB was collected from books in the fields of medicine and Buddhism through manual entry. Given the large number of non-Tibetan characters found on the internet, such as English characters and punctuations that are impractical for subsequent downstream tasks, these characters were uniformly converted to 'N', ensuring that the structure of the Tibetan text was not disrupted.

According to Su et al. (2020) [20], a word is the smallest unit in the Tibetan language that carries semantic meaning, while basic components and syllables lack linguistic information. Phrases and sentences, though containing more information than words, can pose challenges such as excessive data size or high complexity (Zhijie and Dorje, 2023) [21]. Therefore, this study ultimately selected a 'word block' as the unit of granularity for content selection. This segmentation was facilitated by a system developed by Karten et al. (2015) [22]. To ensure textual diversity and coverage, a Tibetan vocabulary construction algorithm based on word frequency is proposed, aiming to develop a broadly applicable Tibetan dataset. The study extracts 10,000 high-frequency Tibetan words from web sources and 7,000 from books. After deduplication, a final corpus of 9,874 frequent Tibetan words is established. These form the dataset's content, as detailed in Table 1.

Table 1. Tibetan Text Content Selection from Variety of Sources Based on Word Frequency and Uniqueness.

No.	Source	Content	Size (MB)	Total	High-Freq	Unique
1	Websites	News	2,058.24	109,819,647	10,000	9,874
2	Books	Medicine, Buddhism	33	2,635,220	7,000	

2.3 Collecting Handwriting Data

2.3.1 Writer Selection

This paper selected Changdu City's Sixth Senior High School, Tibet University, and the Gannan Tibetan Autonomous Prefecture Tibetan Calligraphers Association as the writing groups for the data collection. These groups included not only students but also calligraphers with a high level of writing proficiency, beneficial for capturing different writing styles and characteristics. Due to the high utility of Khyug-yig style in Tibetan script (Yang, 1990) [23], its use was widespread in domains such as education, historical archives, and medical prescriptions. Therefore, the first two groups were students within the Tibet Autonomous Region, from Changdu City's Sixth Senior High School and the Literature Department of Tibet University. These institutions offer Tibetan language courses with standard textbooks, with both

teachers and students being highly skilled in Khyug-yig writing. The third group came from the Gannan Tibetan Autonomous Prefecture. Given the profound calligraphy history of the region, eight candidates were singled out from a local calligraphic association, including the prominent calligrapher Maolanmu. The selection of the aforementioned writer groups aims to facilitate data diversity with various individual characteristics of Khyug-yig style. The data can be further validated through monitored writing and labeling procedures, thereby supporting downstream model advancement in relevant research. Details about each writer group are shown in Table 2.

Table 2. Comprehensive Overview of Data Collection Involving Syllabic Content Types Across Educational Institutions and Association.

No.	Collection Targets	Grade	Tibetan Word Categories	Participants	Num
1	Changdu City’s Sixth Senior High School	Third Year	Monosyllabic, Bisyllabic	45	8,534
2	The Literature Department of Tibet University	Second Year	Multisyllabic	5	178
		Third Year	Multisyllabic	5	177
3	Gannan Tibetan Calligraphers Association	-	Ancient	8	985
Total				63	9,874

2.3.2 Data Collection

After acquiring 9,874 Tibetan words from several web pages and books, the study developed a writing form with essential details that encompassed writer identifications, educational background, and word content with formatting grids. This form ensured the quality of writing and expedited the process of image segmentation and text annotation in subsequent stages.

Furthermore, the writing assignments were allocated to three distinct categories based on writing difficulties and word complexity. Norboo (1976) [24] defined Tibetan words into modern and ancient groups by origins and structures. The first two categories belonged to contemporary Tibetan, with the first comprising monosyllabic or bisyllabic words, while the second included multisyllabic words. Because of the skills required for handwriting, the tasks were appropriately assigned to high school and university students. The third ancient category was entrusted to Khyug-yig style professionals from the calligraphic association due to the complex word structures. As a result, 63 writing forms were collected. Table 2 also details the data collection process schedule.

2.3.3 Data Processing

After collecting handwriting forms from multiple calligraphers, we performed the following steps that process each handwriting example into a more consistent format with ground-truth labels from

Tibetan handwriting experts, ensuring a better quality and delivery of Tibetan Handwriting Khyug-yig Dataset.

Scanning and Resizing. At the first stage, each handwriting form was scanned into a high-resolution image with subsequent grid cropping. This approach converted the Khyug-yig handwriting into word-sized image patches that were uniform in both data size and type.

Image Preprocessing. Each patched image was converted to grayscale to reduce data size, as all handwriting samples were created using black pens on white paper. We next rescaled their pixel values from 0 - 255 to 0 - 1. This normalization technique helps avoid excessively large training examples, which might impede the training process and slow down convergence for many learning algorithms in Tibetan fields. After rescaling, some noises, such as blurred handwriting and ink taints were discovered during the investigation of the image samples and pixel distributions. To remove these noises, we kept only pixels with values ranging from 0 to 0.7 to retain the major features of handwritings, while setting the remaining pixels to 0 (white) to make the object clearer. Figure 3 demonstrates the comparison between the original and preprocessed handwriting samples.

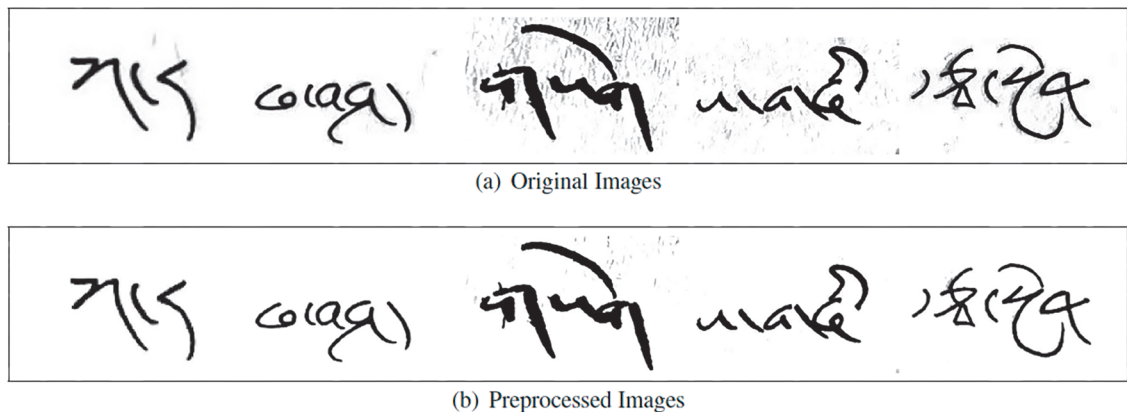


Figure 3. Comparison of original and preprocessed handwriting samples. The original images (a) show the handwritten samples that have been scanned and cropped, including blurred pen strokes and ink taints. The preprocessed images (b) illustrate the results after grayscale conversion, normalization, and noise removal, improving clarity and highlighting the major features of the handwriting.

Grouping and Labeling. After preprocessing, each image was grouped by their writers' names and sorted into a local database. Through organising necessary information such as data sources and correct writer identities, users can efficiently retrieve desired data efficiently, ensuring dataset comprehensiveness, as well as potentially benefiting future work. Lastly, all images were annotated with ground-truth labels provided by experts in Tibetan Khyug-yig handwriting, to further improve the quality and trustworthiness of the constructed dataset. Table 3 shows an example of our Tibetan handwritten Khyug-yig script dataset.

Table 3. Detailed Information of Collected Tibetan Word Data Including Quantity, Data Size, and Writer Styles.

Tibetan Word Categories	Number of images	Size (MB)	Writer Count
Monosyllabic, Bisyllabic	8,534	862	45
Multisyllabic	355	62.22	10
Ancient	985	99.5	8
Total	9,874	1,023.72	63

3. RESULTS ANALYSIS

3.1 Dataset Statistics

The Tibetan handwritten Khyug-yig style exhibits variations in writing based on word structure. This study categorizes words in images to analyze the Khyug-yig dataset. We partitioned the dataset into ancient Tibetan words and modern Tibetan words, where the modern Tibetan words include monosyllabic, bisyllabic, and multisyllabic words, ensuring diversity and comprehensiveness. Since the number of characters in a syllable can affect features of the handwriting such as font shape, stroke thickness, and writing slant, we further divided monosyllabic Tibetan words by character count, as demonstrated in the subsequent Figure 4. The detailed quantitative distribution of each type of Tibetan words is shown in Table 4.

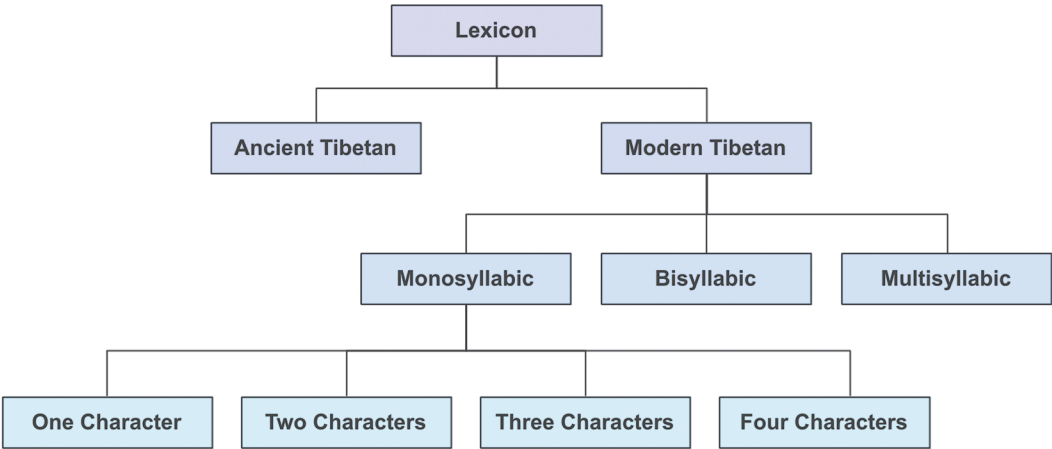


Figure 4. Categorization of Tibetan Words in Dataset Based on Structure, Syllable Types and Character Count.

Table 4. Categorization and Distribution of Modern and Ancient Tibetan Words by Syllable and Character Count.

Word Type	Total	Total(%)	Category	Num	Num(%)	Character Count	Num	Num(%)
Modern Tibetan	8,889	90.03%	Monosyllabic	2,995	33.69%	1	329	10.99%
						2	1,376	45.94%
						3	1,078	35.99%
						4	212	7.08%
			Bisyllabic	5,539	62.32%	-	-	-
			Multisyllabic	355	3.99%	-	-	-
Ancient Tibetan	985	9.97%	-	-	-	-	-	-

In the dataset, modern Tibetan words, such as ‘Economy’ and ‘You’, are the predominant categories, representing 90.03% of Khyug-yig words, while commonly used ancient Tibetan words account for only 9.97%. In addition, we categorize modern Tibetan words in the dataset into monosyllabic, bisyllabic, and multisyllabic types, with proportions of 33.69%, 62.32%, and 3.99%, respectively. Figure 5 depicts examples of handwriting Khyug-yig words across multi-categories. Figure 5 illustrates the distribution of structure types among modern Tibetan words.

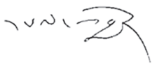





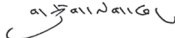





Categories	Examples of Handwriting Khyug-yig Words		
Modern	 Economy	 Stone	 You
Modern Bisyllabic	 Snow	 Flower	 Longevity
Modern Multisyllabic	 Library	 Capitalism	 Proletariat
Ancient	 Magic	 Hong(Exclamation)	 Human

Figure 5. Distribution of Modern Tibetan Khyug-yig Handwritten Words by Syllable Type and Character Count.

As shown in Figures 5 and 6, bisyllabic words such as 'Snow', 'Flower', and 'Longevity' are predominant, with 5539 occurrences, whereas multisyllabic words including 'Library', 'Capitalism', and 'Proletariat' are less common. The dataset further classifies monosyllabic words by number of characters, accounting for 10.99%, 45.94%, 35.99%, and 7.08% of the dataset, respectively. Contributed by various calligraphers, the dataset encompasses a broad spectrum of writing features. Notably, the number of characters in a syllable can result in different writing styles in aspects such as font shape, stroke thickness, and calligraphic slant, with examples shown in Figure 7. This Tibetan Handwritten Khyug-yig Dataset might be able to cover a vast array of Khyug-yig writing features, offering diversity and low redundancy, thus supporting research in Tibetan language recognition.

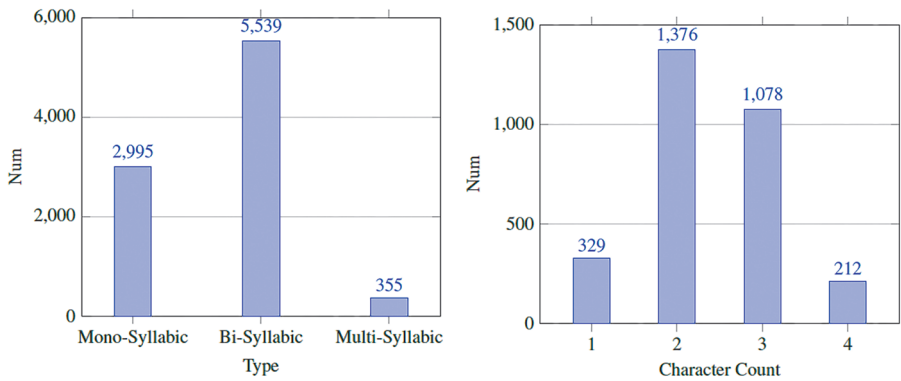


Figure 6. Examples of the same characters in different writing styles resulting from varying numbers of characters within multiple syllabic compositions.

Characters	Number of Characters in a Tibetan Syllable			
	1	2	3	4
ག				
མ				
ས				

Figure 7. Examples of the same characters in different writing styles resulting from varying numbers of characters within multiple syllabic compositions.

3.2 Comparison with Existing Tibetan Handwriting Datasets

To evaluate the strengths and weaknesses of our constructed Tibetan Handwriting Khyug-yig Dataset, we compare it with the existing Tibetan handwriting datasets mentioned in Section 2.3 across several key dimensions: basic units of Tibetan, calligraphic styles, domains, writer count, and data size.

As shown in Table 5, TibetanMNIST [18] contains only numerals in Umê, and Tibetan Handwritten Consonants Dataset [19] includes only simple consonants in Uchen. In both cases, individual numerals or consonants do not carry specific meanings. However, Tibetan Handwriting Khyug-yig Dataset encompasses a total of 9,874 Tibetan numerals, consonants, and words, covering multiple domains such as news, medicine, and Buddhism. Unlike Umê and Uchen, Khyug-yig featured in our dataset is the most intricate, abstract, and commonly used calligraphic style in the daily lives of Tibetan people. Its unique feature, where consonants and syllables are joined, offers significant advantages in handling downstream tasks related to Tibetan cursive script, thereby supporting further research in Tibetan fields. Despite the smaller size of our dataset compared to the other two due to scarce existing resources and high costs, the expansion of the Khyug-yig dataset is in progress.

Table 5. Examples of Tibetan handwriting Khyug-yig words across multi-categories.

Dataset	Unit			Calligraphic Style	Domains	Number of Writers	Number of Images
	Numeric	Consonant	Word				
TibetanMNIST	0-9	-	-	Umê	-	-	17,768
Tibetan Handwritten Consonants Dataset	-	30 basic consonants	-	Uchen	-	150	77,636
Tibetan Handwriting Khyug-yig Dataset	9,874 unique words, including numerals and consonants			Khyug-yig	News, Medicine, Buddhism	63	9,874 (in progress)

4. CONCLUSION

In this study, we present a methodology for constructing the Tibetan Khyug-yig handwritten dataset, concentrating on the pre-collection of textual content and emphasizing lexicon construction and categorization. The extant Khyug-yig dataset includes 9,874 words, noted for their calligraphic diversity and precise labeling. Addressing issues of data scarcity and content repetition, this dataset plays a crucial role in enhancing Tibetan handwriting recognition models. Future efforts aim to expand the dataset, involving expert contributors to yield approximately 20,000 high-quality labeled images, primarily of high-frequency Tibetan words. Such an expansion is instrumental for evolving sophisticated algorithms and applications in Tibetan language technology, contributing significantly to digitization efforts in areas like Tibetan medicine and historical documentation.

Author Contributions

Dorje Tashi (1336786645@qq.com) was pivotal in conceptualizing the study and designing the research framework. He brought extensive knowledge of Tibetan calligraphy, and played a critical role in data collection. Tianying Sheng (tshe5072@uni.sydney.edu.au) and Bingtian Chen (edu.bt.chen@gmail.com) were instrumental in conducting the research, processing the data, and contributing to the majority of the writing and revision processes of the manuscript. Renzeng Duojie (1046447973@qq.com) and Rinchen Dongrub (651130607@qq.com) provided expertise in Tibetan literature and were co-responsible for the collection and analysis of the data. Yongbin Yu (ybyu@uestc.edu.cn) provided expert guidance on the formulation of the conceptualization and the structure of the research framework. Xiangxiang Wang (xxwang@uestc.edu.cn) participated in the implementation of the approach, and provided technical support. Nyima Tashi (nmzx@utibet.edu.cn) offered professional guidance in research design, data analysis, and significant direction during the writing phases, shaping the manuscript's scholarly narrative. All authors have discussed the results and meticulously reviewed the final manuscript.

Acknowledgements

This work was supported in part by the National Science and Technology Major Project under Grant 2022ZD0116100, in part by the National Natural Science Foundation of China under Grant 62276055, in part by the Sichuan Science and Technology Program under Grants 23ZDYF0755,24NSFSC5679. We would like to extend our sincere gratitude to the authors from Changdu City's Sixth Senior High School, the Literature Department of Tibet University, and Gannan Tibetan Calligraphers Association, for their participation in the creation of Tibetan Handwriting Khyug-yig Dataset.

Data Availability Statement

The dataset generated and analyzed during the current study is available in the GitHub repository: <https://github.com/13209413223/Tibetan-Handwriting-Dataset>. Please note that the construction work of our dataset is still in progress, and more data will be gradually released in the future.

References

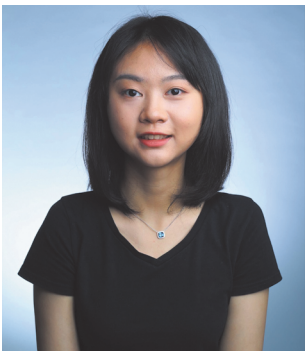
- [1] Bao, W., Dong, J., Xu, Y., Yang, Y., Qi, X.: Exploring Attentive Siamese LSTM for Low-Resource Text Plagiarism Detection. *Data Intelligence*, 1–15 (2023)
- [2] An, B.: Prompt-based for low-resource Tibetan text classification. *ACM Transactions on Asian and Low-Resource Language Information Processing* 22(8), 1–13 (2023)
- [3] Quenzer, J., Bondarev, D., Sobisch, J.: Towards a Tibetan Palaeography: Developing a Typography of Writing Styles in Early Tibet. In: *Manuscript Cultures: Mapping the Field*, pp. 299–441. De Gruyter, Berlin (2014)
- [4] Yang, X.: Research on Text Recognition of Dunhuang Tibetan Ancient Books. MSc dissertation, Tibet University (2023). Available at: <https://link.cnki.net/doi/10.27735/d.cnki.gxzdxd.2023.000423>. Accessed 10 Mar 2024

- [5] Dhondup, R., Tsering, T., Tashi, N.: Study on a Synthesis Method for Training Data of Ancient Tibetan Book Character Recognition. *Plateau Science Research* 3, 84–91 (2021)
- [6] Tong, P., Long, B., Yong, C.: Tibetan Umei Scene Character Recognition Based on Deep Learning. *China Computer and Communication* 35(4), 91–93 (2023)
- [7] Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F.: ImageNet: A Large-Scale Hierarchical Image Database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009)
- [8] Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
- [9] Von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 319–326 (2004)
- [10] Rosch, E.: Principles of categorization. In: *Cognition and Categorization*, pp. 27–48. Lawrence Erlbaum Associates, Mahwah (1978)
- [11] LeCun, Y., Cortes, C., Burges, C.: THE MNIST DATABASE of handwritten digits. Available at <http://yann.lecun.com/exdb/mnist> (1994). Accessed 16 Dec 2023
- [12] Cohen, G., Afshar, S., Tapson, J., Van Schaik, A.: EMNIST: Extending MNIST to Handwritten Letters. In: 2017 International Joint Conference on Neural Networks (IJCNN), pp. 2921–2926 (2017)
- [13] Grother, P.J.: NIST special database 19. Handprinted Forms and Characters Database, National Institute of Standards and Technology 10, 69 (1995)
- [14] Marti, U.-V., Bunke, H.: The IAM-database: An English Sentence Database for Offline Handwriting Recognition. *International Journal on Document Analysis and Recognition* 5(1), 39–46 (2002)
- [15] Johansson, S., Leech, G.N., Goodluck, H.: Manual of Information to Accompany the Lancaster-Oslo: Bergen Corpus of British English, for Use with Digital Computers. Department of English, University of Oslo (1978)
- [16] Wang, D., Liu, C., Yu, J., Zhou, X.: CASIA-OLHWDB1: A Database of Online Handwritten Chinese Characters. In: 2009 10th International Conference on Document Analysis and Recognition, pp. 1206–1210 (2009)
- [17] Zhou, S., Chen, Q., Wang, X.: HIT-OR3C: An Opening Recognition Corpus for Chinese Characters. In: Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, pp. 223–230 (2010)
- [18] bat67.: TibetanMNIST. Available at <https://github.com/bat67/TibetanMNIST> (2019). Accessed 16 Dec 2023
- [19] Crxm.: Tibetan Handwriting Consonants Dataset. Available at <https://www.heywhale.com/mw/dataset/5eb0d52f366f4d002d756691> (2020). Accessed 24 Dec 2023
- [20] Su, H., Lamu, S., Tashi, N., Nuo, Q.: Research on Tibetan Text Classification Based on MLP and SepCNN Neural Network Model. *Computer Engineering and Software* 41(12), 11–17 (2020)
- [21] Zhijie, C., Tashi, D.: Feature Primitives Selection for Tibetan Text Classification. *Journal of Chinese Information Processing* 37(1), 64–70 (2023)
- [22] Karten, L., Yang, Y., Zhao, X.: Tibetan Automatic Word Segmentation Based on Conditional Random Fields and Knowledge Fusion. *Journal of Chinese Information Processing* 29(6), 213–219 (2015)
- [23] Yang, Z.: A Brief Introduction to Tibetan Calligraphy Art. *Journal of Southwest Minzu University (Humanities and Social Sciences Edition)* 2, 115–116 (1990)
- [24] Norboo, S.: A Short History of Tibetan Translated Literature. *The Tibet Journal* 1(3/4), 81–84 (1976)

AUTHOR BIOGRAPHY



Dorje Tashi is a PhD candidate at the School of Information Science and Technology, Tibet University. His academic research primarily focuses on the fields of Tibetan computational linguistics and pattern recognition. ORCID: 0009-0002-8690-8185



Tianying Sheng is a first-year postgraduate student specializing in Data Science and AI at the School of Computer Science, University of Sydney. She is interested in scalable database management and statistical analytics. As part of her study, she has worked as a teaching assistant to deliver fundamental and advanced database classes for undergraduates. ORCID: 0009-0001-2443-0406



Bingtian Chen is currently pursuing a Master's degree in Software Engineering at the School of Information and Software Engineering, University of Electronic Science and Technology of China (UESTC). His research interests include computer graphics, databases, and web development. He has been involved in several projects that emphasize practical application of software engineering techniques and has achieved progress in optimizing high-performance database subqueries. ORCID: 0009-0009-1459-1888



Renzeng Duojie, holding a PhD, is a lecturer at the School of Information Science and Technology, Tibet University. His primary research interests focus on Tibetan computational linguistics and speech processing.
ORCID: 0009-0008-3063-2443



Rinchen Dongrub, who has earned a PhD, serves as a lecturer at the School of Information Science and Technology, Tibet University. His research is deeply rooted in Tibetan computational linguistics and pattern recognition.
ORCID: 0009-0000-7608-8078



Yongbin Yu was born in Sichuan Province, China, in 1975. He received his PhD degree from the University of Electronic Science and Technology of China (UESTC), Chengdu, in 2008. He visited the University of Michigan-Ann Arbor in 2013 and the University of California-Santa Barbara in 2016. He won the first prize of the science and technology award of the Tibet Autonomous Region in 2018. He worked as a ‘guest’ deputy director in the department of big data industry, Sichuan Provincial Economic and Information Commission in 2018. Currently, he is an associate professor in the School of Information and Software Engineering, UESTC. His research interests include memristor-based neural network, swarm intelligence, natural language processing and big data.
ORCID: 0000-0001-6022-7504



Xiangxiang Wang was born in Henan Province, China. He received his PhD degree from the University of Electronic Science and Technology of China (UESTC), Chengdu, in 2023. From June 2021 to June 2022, he was a joint PhD Student with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, sponsored by the Academic Exchange Special Fund in UESTC for Overseas Training. Currently, he is a lecturer in the School of Information and Software Engineering, UESTC. His current research interests include memristive neural networks, complex neural networks, impulsive control, natural language processing and synchronization analysis.

ORCID: 0000-0001-9341-1068



Nyima Tashi, PhD in Engineering and an academican of the Chinese Academy of Engineering, is the dean, professor, and PhD supervisor at the School of Information Science and Technology, Tibet University. His research work centers on Tibetan information technology, computational linguistics, and Tibetan information systems.

ORCID: 0000-0001-9288-6600