Data Collection and Preparation Process

3.1 Overview

The development of an effective OCR system for Sylheti Nagri script required a comprehensive and multi-faceted approach to data collection and preparation. Given the limited availability of digitized Sylheti Nagri texts and the absence of existing annotated datasets for this script, we employed three distinct methodologies to create a robust training corpus totaling **151,402 samples** (**124,366 training** + **27,036 evaluation**).

3.2 Data Collection Methodologies

3.2.1 Manual Annotation from Historical Texts (Method 1)

The first phase involved collecting authentic Sylheti Nagri texts from historical sources to ensure linguistic accuracy and preserve the traditional characteristics of the script. We selected six seminal works representing different genres and time periods:

- **সাত কন্যার বাখান** by সৈয়দ শাহনূর (Folk literature)
- **ছদছি মছলা** by আব্দুল করিম (Religious text)
- **দেশ চরিত** by ছৈয়দুর রহমান (Historical narrative)
- **সোনাভানের পুথি** by মুন্সী আব্দুল করিম (Traditional poetry)
- **খবর নিশান** by ফকির ভেলা শাহ (Spiritual literature)
- **শ্রীহট্টিয়া ছড়া** by মীর লিয়াকত আলী (Regional poetry)

Data Processing Pipeline:

1. **Image Acquisition**: High-resolution photographs were taken of original book pages using standardized lighting conditions
2. **Preprocessing**: Images were enhanced using CamScanner application to correct perspective distortion, improve contrast, and normalize lighting
3. **Word Segmentation**: Individual words were automatically cropped using a custom Python implementation leveraging PaddleDBNet for text detection
4. **Manual Annotation**: Each cropped word image was manually labeled by examining the original text, with annotations stored in structured Excel files containing image-label pairs

This method yielded **10,503 training samples** and **4,158 evaluation samples** (total: **14,661 samples**), ensuring high-quality ground truth data with authentic script variations.

3.2.2 Web Scraping and Synthetic Data Generation (Method 2)

To address the scarcity of digital Sylheti Nagri content, we implemented a systematic web scraping approach targeting online repositories:

Source Identification:

- Sylheti Wikipedia (https://syl.wikipedia.org/)
- Sylheti Kitab digital library (https://www.sylhetikitab.org/en/read)
- Tourat Shorif in Sylheti Nagri script

Data Extraction and Processing:

1. **Web Scraping**: Automated extraction of Sylheti Nagri text using custom Python scripts
2. **Text Curation**: Extracted content was manually reviewed and organized in Google Docs for quality assurance
3. **Synthetic Image Generation**: Individual words were converted to images using PIL (Python Imaging Library) with authentic Sylheti Nagri fonts:
   - **Primary Font**: Noto Sans Syloti Nagri (Google Fonts)
   - **Secondary Font**: Surma-Regular.ttf (custom Sylheti font)

Image Generation Parameters:

- **Font Selection**: Two high-quality Sylheti Nagri fonts were employed:
  - **Noto Sans Syloti Nagri**: Modern Unicode-compliant font (Google Fonts standard)
  - **Surma-Regular.ttf**: Traditional Sylheti typeface for stylistic diversity
- **Rendering Specifications**:

```
python

font_size = 24 pixels
image_background = white (RGB: 255,255,255)
text_color = black (RGB: 0,0,0)
padding = 10 pixels (horizontal and vertical)
```

Font-Specific Generation Results:

- **Noto Sans Font Dataset**: 78,977 training samples + 12,475 evaluation samples = **91,452 samples**
- **Surma Font Dataset**: 34,886 training samples + 10,403 evaluation samples = **45,289 samples**

Combined, this methodology produced **136,741 samples**, significantly expanding the dataset while maintaining typographic authenticity and providing font variation for improved model generalization.

3.2.3 Data Augmentation for Robustness (Method 3)

To enhance model generalization and simulate real-world document variations, we applied sophisticated augmentation techniques to the existing annotated datasets.

Augmentation Pipeline: The augmentation strategy was designed to replicate authentic document aging and scanning artifacts commonly found in historical manuscripts:

- **Geometric Transformations**:
  - Rotation: ±2° (p=0.4)
  - Affine scaling: 0.99-1.01 (p=0.3)
  - Translation: 0.5-1% of image dimensions (p=0.3)
  - Elastic deformation: α=0.5, σ=30 (p=0.2)
- **Photometric Variations**:
  - Brightness adjustment: -15% to -5% (p=0.3)
  - Contrast enhancement: ±10% (p=0.3)
  - Gaussian blur: 1-2 pixel radius (p=0.2)
  - Motion blur: 2 pixel limit (p=0.1)
- **Noise and Degradation**:
  - ISO noise: 0.5-2% color shift, 5-15% intensity (p=0.2)
  - Multiplicative noise: 98-102% (p=0.2)
  - Sepia tone effect (p=0.3)
  - JPEG compression: 50-70% quality (p=0.1)

This augmentation process was applied to existing datasets to enhance model robustness, generating additional variations that improve generalization to document quality variations. The augmentation contributed to the overall sample count by creating enhanced versions of the base datasets.

3.3 Dataset Composition and Statistics

The final compiled dataset consists of:

Data Source	Training Samples	Evaluation Samples	Total Samples	Percentage
Historical Texts (Manual)	10,503	4,158	14,661	9.7%
Noto Sans Font (Synthetic)	78,977	12,475	91,452	60.4%
Surma Font (Synthetic)	34,886	10,403	45,289	29.9%
Total	**124,366**	**27,036**	**151,402**	**100%**

Dataset Split Ratio: Training (82.1%) : Evaluation (17.9%)

3.4 Quality Assurance and Validation

Manual Verification: A stratified sample from each methodology was manually reviewed for annotation accuracy, with particular attention to:

- Historical text annotation consistency with original sources
- Synthetic data rendering quality and Unicode compliance
- Font variation impact on character recognition

Character Distribution Analysis: The dataset encompasses the complete Sylheti Nagri Unicode range (U+A800–U+A82F), ensuring comprehensive coverage of:

- 44 base characters
- 4 independent vowels
- 7 dependent vowel signs
- Various conjunct forms and diacritical marks

Font Diversity Impact: The inclusion of both Noto Sans and Surma fonts provides:

- **Typographic Variation**: Different character stroke weights and stylistic approaches
- **Rendering Differences**: Variations in character spacing, ligature formation, and diacritical mark positioning
- **Real-world Applicability**: Coverage of both modern digital typography and traditional manuscript styles

Word Length Distribution:

- Average word length: 4.2 characters
- Range: 1-15 characters
- Most common length: 3-5 characters (approximately 68% of dataset)

This multi-faceted approach to data collection and preparation ensures both the authenticity and diversity necessary for training a robust Sylheti Nagri OCR system while addressing the unique challenges posed by this historically significant but computationally underrepresented script.