


# Enhanced HTR Accuracy for Tibetan Historical Texts - Optimising Image Pre-processing for Improved Transcription Quality

Christina Sabbagh

(SOAS, University of London)

 Automatic transcription technology, or handwritten text recognition (HTR), unlocks new opportunities for analysing historical texts by transforming document images into machine-readable formats. This paper explores the development and evaluation of an image pre-processing pipeline to improve transcription accuracy for Tibetan historical newspapers. Images of historical texts often suffer from degradations introduced throughout their lifecycle, negatively affecting HTR accuracy. Additionally, the variability in image quality across archives poses challenges to creating a universally applicable pre-processing pipeline. This study compares three pre-processing pipelines, ultimately revealing a dynamic approach that could adapt to varying document conditions and that resulted in the highest transcription accuracy. This method offers a replicable solution for future research. We have also made the source code publicly available to support further exploration.

## 1 Introduction

Historical document preservation relies increasingly on digital transformation technologies. For Tibetan historical materials, this process confronts significant challenges: documents often suffer from

deterioration, including fading ink, paper damage, and inconsistent image quality. Such degradations impede handwritten text recognition (HTR),<sup>1</sup> which is essential for converting physical documents into searchable, analysable digital resources.

Current HTR technologies struggle with historical documents, particularly those with complex visual characteristics such as Tibetan newspapers in cases where the image quality is impaired. Manual image enhancement is time-consuming and impractical for large collections, necessitating automated solutions. Our research developed a pre-processing pipeline designed to improve image quality and the resulting transcription accuracy for these challenging historical documents.

This study addresses three primary questions:

- (1) Can targeted image pre-processing techniques effectively improve HTR transcription accuracy for degraded Tibetan historical documents such as newspapers?
- (2) How do different pre-processing approaches (traditional, deep learning, and hybrid) compare in addressing image quality challenges?
- (3) What methodology provides the most reliable and adaptable solution for transcribing historical Tibetan document images?

We compared three distinct image processing approaches: a traditional binarisation method, a deep learning-based technique, and a novel hybrid approach that dynamically selects the most appropriate method based on image quality. By evaluating these techniques, we

---

<sup>1</sup> While optical character recognition (OCR) refers to the translation of printed documents into machine-readable text, handwritten text recognition (HTR) is a similar process designed to handle the challenges of recognising text written in variable or inconsistent fonts, such as those created by handwriting or the differing fonts used across newspapers (Nockels *et al.* 2024: 149-150). HTR systems must account for these variations in style, size, and slant, which makes the process more complex than OCR.

aim to provide researchers with a robust, adaptable tool for digital document preservation.

By transforming deteriorating and rare documents into digital resources, this research offers an original approach to historical preservation by increasing the possibilities for large-scale analysis, broad accessibility, and long-term conservation of cultural heritage.

## 2 *Background*

### 2.1 *Dataset characteristics*

Experiments were conducted on a dataset provided by the “Divergent Discourses” project which applies digital philology methodologies to investigate the complex narratives of Tibetan history and identity emerging after the 1950 annexation of Tibet by the Chinese People’s Liberation Army. Central to this research is a collection of historical newspaper images curated from multiple archives, with handwritten text recognition (HTR) serving as a critical computational tool for transforming the document images into searchable digital resources.

As detailed by Erhard (2025) in this special issue, the Divergent Discourses Corpus of Tibetan-language newspapers currently comprises 16,718 pages from 16 newspapers, sourced from eight private collections and library archives across India, the United States, the United Kingdom, and Europe. Of these, 7,341 images are predominantly high-quality scans of original newspapers. In contrast, the majority – 9,377 images – were provided by the Staatsbibliothek zu Berlin (SB) and were digitised from microfilm. The microfilm, likely created in the late 1990s, was published by the China National Microforms Import & Export Corporation, Beijing, in the 2000s and presents several challenges. It includes missing and duplicated pages, as well as underexposed images with dark patches. Additionally, the collection preserved on microfilm shows signs of wear and water damage which poses significant difficulties for our use-case.

Drawing on the framework by Alaei *et al.* (2023), we analysed the document degradations visible within the dataset across three stages of a document's life cycle (Fig 1):

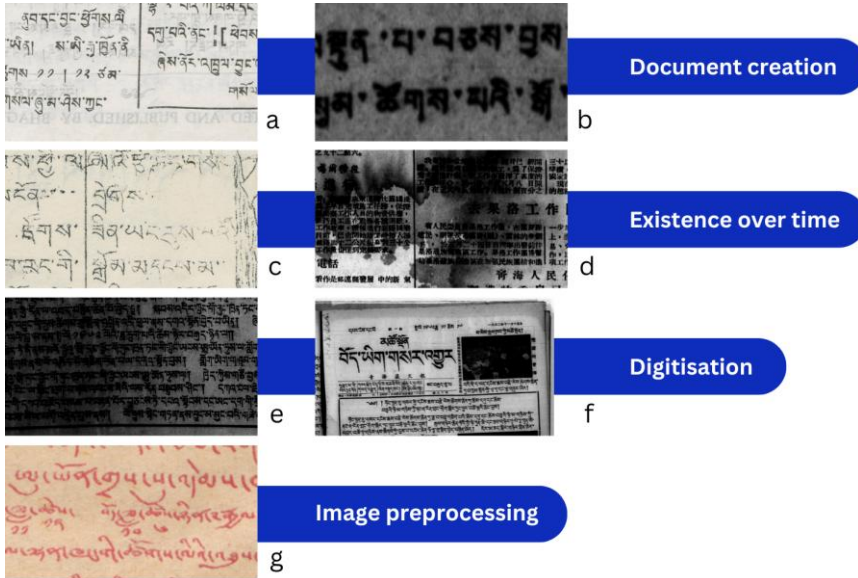


Figure 1 Visual examples of image degradation introduced at different stages within the document life cycle: a) bleed-through, b) blotchiness, c) fading, d) staining, e) uneven illumination, f) skewing, g) red text.<sup>2</sup>

- (1) Document Creation:
  - Bleed-through
  - Blotchy text, likely from low-quality paper or ink
- (2) Existence Through Time:
  - Text fading
  - Paper staining
- (3) Digitisation:
  - Uneven illumination during capture
  - Inconsistent page orientation
  - Compression and quality variations

<sup>2</sup> a) Tibetan Freedom, Nov. 1, 1965 p3; b) Tibet Daily, Jan. 1 1962 p1; c) Defend Tibet's Freedom, Aug. 20 1963 p3; d) Qinghai Tibetan News, Jun. 14 1955 p4; e) Qinghai Tibetan News, Jul. 5 1951 p2; f) Qinghai Tibetan News, Nov. 15 1952 p1; g) News in Brief, Dec. 1 1953 p1.

Red-text documents pose an additional challenge as they require distinct treatment during pre-processing compared to those containing black text. Failure to account for these chromatic differences generally resulted in less accurate HTR transcriptions, as traditional pre-processing approaches often reduced legibility of red text rather than improving it.

These cumulative degradations across the document life cycle obscure the textual features necessary for machine recognition, rendering standard HTR processes insufficient for reliable transcription.

## 2.2 *Image quality variation and representation bias*

Historical document collections inherently contain variability in quality arising from diverse archival sources, each reflecting distinct socioeconomic, geographical, and preservation contexts. This can result in variable HTR quality, introducing potential representation bias that could distort computational and historical analysis.

Ehrmann *et al.* (2023) illustrate this risk: a detected drop in word frequency during a historical period such as a war might not represent a genuine linguistic shift but could instead result from compromised paper quality. Newspapers from lower socio-economic strata may have been produced using lower-quality materials, increasing the likelihood of degradation over time and impacting HTR accuracy (Beelen *et al.* 2023). Such risks underscore the importance of effective image pre-processing.

Further, varying institutional digitisation strategies (Coutts 2016) can result in image quality differences. Research also suggests a correlation between digitisation quality and available financial resources (Smith & Cordell 2018). Institutions with limited economic and technical capabilities may produce lower-quality digital representations across different stages of the document life cycle. This variation in digitisation strategies can result in disproportionately poor-quality images within certain regions, potentially skewing research insights.

Image pre-processing techniques offer a methodological intervention, standardising transcription accuracy across heterogeneous document collections. By addressing image degradation introduced at each stage of a document's life cycle, researchers can minimise bias and enhance the reliability of computational and historical analyses.

This approach transforms technical challenges into an opportunity for more nuanced, comprehensive historical research, ensuring that marginal or less-preserved documents receive equal scholarly attention.

### 2.3 *Image binarisation*

Image binarisation is a frequently used pre-processing technique in digital document analysis that transforms complex images into black-and-white representations. By converting each pixel to either foreground (black) or background (white), binarisation helps isolate text and improve its machine-readability, particularly for historical documents with varying image qualities.

Researchers have developed three primary approaches to binarisation:

- **Uniform Thresholding:** Applies the same thresholding rule across the entire image, exemplified by the method developed by Otsu (1975).
- **Locally Adaptive Thresholding:** Tailors the threshold for each pixel based on local pixel neighbourhoods, making it particularly effective for images with uneven illumination. Sauvola (2000) and Niblack (1986) thresholding are examples of this approach.
- **Deep Learning-Based Methods:** Exemplified by the Berlin State Library's (SBB) binarisation model (Rezanezhad 2023), these methods use neural networks to make pixel-wise decisions based on both global and local image characteristics simultaneously.

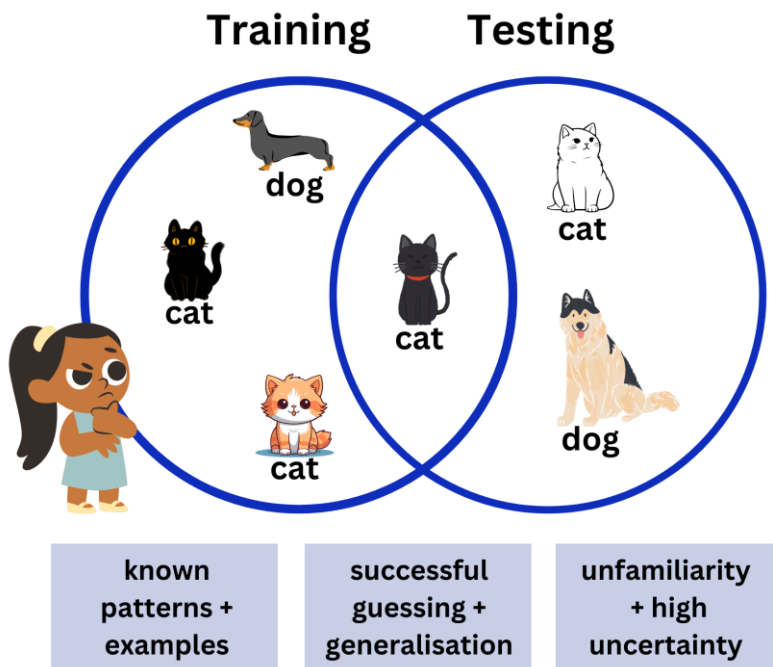


Figure 2 *A deep learning model learns to recognise patterns in data through training, much like how a child learns by studying examples. For instance, a parent may point to animals and label them as ‘dog’ or ‘cat.’ If the child only encounters small dogs and black or ginger cats during this learning phase, they become skilled at identifying those specific examples. However, in the real world, dogs vary in size, and cats can have different colours. When the child encounters a black cat they haven’t seen before, they can still correctly identify it. In contrast, encountering a large dog or a white cat may lead to misclassification. The more diverse the examples the child is exposed to, the better they become at generalising and accurately identifying unfamiliar animals.*

Different studies have yielded conflicting conclusions about which of these approaches best enhances HTR accuracy on document images. Some researchers found Otsu’s method superior for historical documents (Gupta *et al.* 2007; Rawat *et al.* 2021; Taş & Müngen 2021), while others demonstrated the effectiveness of locally adaptive techniques such as “Niblack Thresholding” (Jacsont & Leblanc 2023). These discrepancies likely arise from variations in underlying HTR models and the heterogeneous nature of document datasets.

More recently developed deep learning approaches, while promising, are not without limitations. These methods rely on training datasets to 'learn' patterns and make decisions, much like how humans learn by studying examples (Fig. 2). For instance, if a deep learning model is trained on a set of high-quality, clear document images, it becomes skilled at processing documents that look similar to those it has seen before. However, this process also means that the model might struggle with documents that look very different from its training examples, such as historical newspapers with stains, faded ink, or unusual layouts.

A common workaround is to use synthetic training data - artificially generated images designed to mimic real documents. While this is helpful for creating large datasets, it does not always prepare the model to handle the messy, unpredictable nature of real-world documents (Zhou *et al.* 2023). In contrast, traditional algorithmic methods do not rely on training data in the same way. Instead, they apply fixed rules and processes, which can make their performance more consistent across diverse document types and conditions.

These traditional methods often perform just as well as – or even better than – deep learning methods for certain tasks (Lins *et al.* 2021). This is because deep learning methods depend on their training datasets, which qualitatively align with the documents researchers plan to process. If the training data does not represent the target documents well, the model may fail to generalise effectively and produce less accurate results.

#### 2.4 *Document image pre-processing challenges*

The complex degradation-based challenges associated with historical document digitisation (outlined in Section 2.1) often require more than a single-step approach. Several research projects have developed tailored pre-processing strategies to address the degradation-based challenges.

For instance, Griffiths (2024) adjusted image sharpness, resolution, and noise levels to enhance a Tibetan manuscript dataset. Luo & van



der Kuijp (2024) implemented treatments including rotation correction, border removal, and contrast enhancement for Tibetan books and manuscripts. Rawat *et al.* (2021) demonstrated a multi-step approach for Garhwali textbook pages, utilising greyscaling, binarisation, morphological operations (described in Section 3.1.2), and skew correction.

However, not all pre-processing techniques are universally applicable. Some steps, such as border removal, may be redundant with advanced HTR tools such as Transkribus,<sup>3</sup> which can detect text regions without image trimming. Moreover, the lack of open-source methodologies has hindered the reproducibility and adaptability of previously used pre-processing pipelines. Finally, Jacsont *et al.* (2023) cautioned that combining multiple pre-processing treatments might produce lower-quality images than applying a single treatment, underscoring the importance of quantitative evaluation of pre-processing methods.

A limitation in current approaches is the assumption of uniform image quality. While researchers acknowledge that image qualities and required treatments can vary considerably within a single collection (Rawat *et al.* 2021), few have proposed systematic methods to address this variability. Our study addresses this gap by developing an adaptive pre-processing method that can accommodate diverse document characteristics.

Practical constraints further complicate the implementation of advanced pre-processing techniques. Many deep learning approaches (Anvari & Athitsos 2021; Zhou *et al.* 2023) require substantial computational resources and technical expertise to employ, making them challenging for projects with limited resources. Consequently, we prioritised methods that were well-established, openly accessible, and compatible with standard computing systems. The only exception was the deep learning-based binarisation approach which can still be

---

<sup>3</sup> Transkribus is a web-based platform which offers tools for the digitisation, text recognition, transcription and searching of historical documents. The Divergent Discourses project has used it to develop handwritten text recognition (HTR) models for the automatic transcription of its historical newspaper corpus.

run on standard computing systems, but more slowly than with sophisticated hardware.

Python libraries<sup>4</sup> such as OpenCV (Bradski 2000) and scikit-image (van der Walt 2014) offer a robust suite of pre-processing tools that are both extensively tested and accessible. By leveraging these resources, we aimed to develop a practical, reproducible methodology for researchers facing technological constraints.

Importantly, the complexity of pre-processing goes beyond mere technical optimisation. For images of historical documents - particularly those in lesser-studied languages - each image represents a potential trove of unique information. Inappropriate pre-processing can degrade image quality (Jacson & Leblanc 2023), or parts of the image, risking the loss of valuable historical insights. In the Divergent Discourses Corpus, many page images are unique, with no alternative copies available if images feature degradations. This highlights the importance of carefully considered and optimised pre-processing techniques.

## 2.5 *Image quality assessment*

Image quality assessment (IQA) is a technique for quantitatively evaluating document image characteristics, addressing the challenge of determining an image's legibility. Traditionally, image quality has been assessed through two primary approaches: Human-performed perceptual evaluation (subjective) and computational ('objective') models quantifying various forms of image degradation.

These assessment methods are broadly categorised into two types: no-reference and full-reference approaches. No-reference models, such as the one we employed, operate without an ideal-quality comparison image—important when working with unique historical

---

<sup>4</sup> A Python 'module' is a single file containing reusable code such as functions. A 'package' is a collection of related modules organised into folders. The term 'library' is generally used to describe collections of modules and packages that provide tools to perform a task, but it can also refer to a single module or package.

documents where perfect originals may not exist. Full-reference methods require an optimal version of the image for direct comparison.

In our research, we used the MANIQA model (Yang *et al.* 2022), a no-reference IQA tool. While initially developed and trained on a dataset of everyday photographs from the KONIQ-10k dataset (Hosu *et al.* 2020)—which predominantly features natural scenes, portraits, and urban environments—we found its quality assessment capabilities to be applicable to historical document analysis.

The KONIQ-10k dataset features images scored by humans based on quality indicators including noise (random colour or brightness variations that obscure details, such as graininess), compression artefacts, blur, exposure issues, and colour-related distortions. As several of these were degradations which appeared to affect HTR quality, the model demonstrated a reasonable degree of reliability in assessing document image readability across our historical Tibetan newspaper collection.

We systematically tested 18 IQA models, ultimately selecting MANIQA for its superior performance in reflecting our own transcription accuracy-based quality assessments (Appendix A).

### 3 *Experimental setup*

The objectives of our research were to evaluate whether targeted image pre-processing could improve handwritten text recognition (HTR) transcription accuracy for our Tibetan historical newspaper collection, and to identify which preprocessing steps improved recognition accuracy most effectively. We developed three pre-processing methods, comparing their performance against a baseline to identify the most effective image treatment strategies.

### 3.1 *Image pre-processing methods*

#### 3.1.1 *Baseline method*

Our baseline approach represented our minimal agreed requirements for document digitisation, including meeting Transkribus upload requirements as of June 2024. We prepared images by converting them to JPEG format, ensuring a minimum of 2500 pixels in at least one dimension, and maintaining a file size under 10 MB. This method simulated the most basic approach researchers might adopt when digitising historical documents, leaving images lightly pre-processed, providing a reference point for evaluating more sophisticated pre-processing techniques.

#### 3.1.2 *Foundational pre-processing pipeline*

We next developed our foundational pre-processing pipeline, deciding which treatments to include. Drawing on cautionary findings of Jacsont & Leblanc (2023) about the potential risks of combining multiple pre-processing treatments, we first isolated and evaluated individual image enhancement techniques.

We explored several pre-processing treatments (Fig. 3):

- **Greyscaling:** This treatment reduces the computational complexity of subsequent steps. To minimise processing time, we converted images to greyscale before testing the 'isolated' effects of each treatment.
- **Unsharp masking:** This method effectively sharpens character edges but simultaneously accentuates non-textual image features such as fold marks and stains. Our experiments revealed that unsharp masking can introduce additional

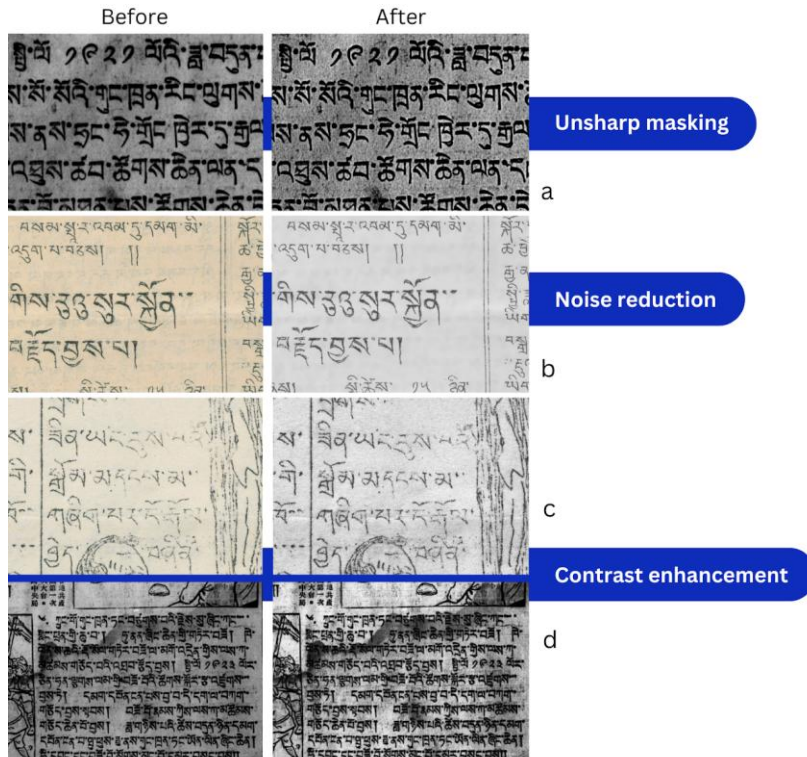


Figure 3 Images before and after having unsharp masking (a), noise reduction (b) and contrast enhancement (c, d) pre-processing treatments applied to them.<sup>5</sup>

speckling and background noise when combined with other treatments. We hypothesised that these artefacts would likely increase HTR transcription errors, as the model might misinterpret the enhanced noise as characters. Consequently, we excluded unsharp masking from our final pre-processing pipeline.

- **Noise reduction:** Using so-called fast non-local means denoising,<sup>6</sup> we effectively minimised digital artifacts such as

<sup>5</sup> a) Qinghai Tibetan News, July 5, 1952, p.4; b) Defend Tibet's Freedom, Aug. 20, 1963, p.8; c) Defend Tibet's Freedom, Aug. 20, 1963, p.3; d) Qinghai Tibetan News, July 5, 1952, p.4.

<sup>6</sup> Fast non-local means denoising aims to remove unwanted noise, such as graininess, from an image while preserving meaningful details. It does this by

page stains, bleed-through, and digitisation-introduced speckling. This technique preserved fine textual details while reducing background noise that could confuse HTR algorithms. We additionally experimented with: gaussian blur and median blur denoising (Bradski 2000). Gaussian and median blur denoising did not make a noticeable difference to the images so we did not progress with these.

- **Contrast enhancement:** To enhance the definition of characters, particularly faint and originally red text, we tested several contrast enhancement techniques. These included methods called: histogram equalisation, adaptive histogram equalisation (CLAHE), and a combination of normalisation (or 'contrast stretching') followed by CLAHE.

Histogram equalisation generally improved the human legibility of darker images but was less effective for lighter images, limiting its utility given the diverse lighting conditions in our dataset. CLAHE proved more effective, enhancing character definition across a wider range of image types. Combining normalisation with CLAHE yielded similar improvements in character definition while further increasing the contrast between originally red text and its background.

This combination marginally enhanced faint text and significantly improved the visibility of red text. However, it also exaggerated existing degradations such as staining and speckling, even after noise reduction. Given these drawbacks, we chose not to include contrast enhancement in our pipeline. Nonetheless, researchers working with less noisy datasets may find this approach beneficial for improving HTR results.

- **Binarisation (thresholding):** This represented our most nuanced intervention. We experimented with multiple thresholding approaches, ultimately finding that Sauvola and deep learning-based SBB binarisation provided the most significant

---

identifying similar patches throughout an image and averaging their values to smooth out the noise.

improvements in human legibility (Fig. 4).



Figure 4 A visual comparison of baseline (lightly treated) images, their counterparts treated with Sauvola and SBB binarisation<sup>7</sup>

<sup>7</sup> Tibet Daily, Jan. 1, 1962, p1 (top); Qinghai Tibetan News, Jun. 21, 1955, p1 (bottom)

We additionally experimented with the Otsu method, mean adaptive binarisation, and Niblack thresholding.

- **Dilation or erosion (morphological operations):** These strategies addressed image characteristics such as faint or blotchy characters (Fig. 5). Dilation increased the definition of faint characters. However, it dilated already blotchy characters, rendering them illegible, and amplified speckling and staining. Erosion intended to narrow blotchy characters often eroded the majority of the text, with no optimal setting that could simultaneously reduce blotchiness while preserving information in faint text. Given these limitations, we did not incorporate dilation or erosion into our final pre-processing pipeline.

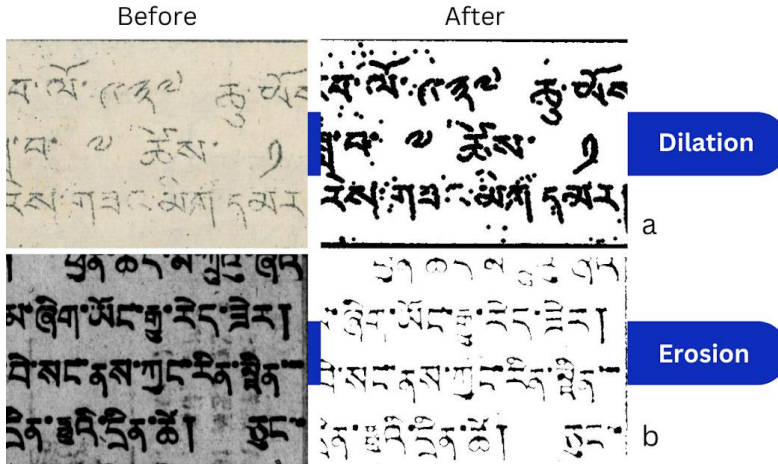


Figure 5 Images before and after dilation and erosion pre-processing operations have been applied to them.<sup>8</sup>

- **Skew correction:** This method addressed rotational irregularities. A projection profiling-based method (Reddy, 2019) appeared to prove most effective for our historical newspaper images. We also experimented with Brunner's (2024) "deskew" Python library, an implementation of the Hough Transform (Panzer, 2017), and the "skew\_correction" Python

<sup>8</sup> a) Defend Tibet's Freedom, Aug. 20, 1963, p.1; b) Qinghai Tibetan News, July 12, 1952, p.3.



library by Bhattarai (2019).

Rawat *et al.* (2021) suggested that super-resolution, used to enhance the resolution of images, was effective for low-resolution images but reduced the HTR accuracy for higher-resolution images. As we did not identify any low-resolution images (approximately 500 x 900 pixels) in our dataset, we did not experiment with super-resolution.

These experiments made the simplifying assumption that the treatments which resulted in the most improved human legibility would also result in the most improvement to machine legibility. Following isolated experiments, we combined the most promising of the above-listed treatments in a pipeline, resulting in the foundational pipeline outlined in Figure 6. The pipeline greyscaled the images, performed fast non-local means denoising, binarisation (either Sauvola binarisation or the deep learning-based SBB binarisation depending on the method), skew correction and compression to ensure the image was 10 MB or less.

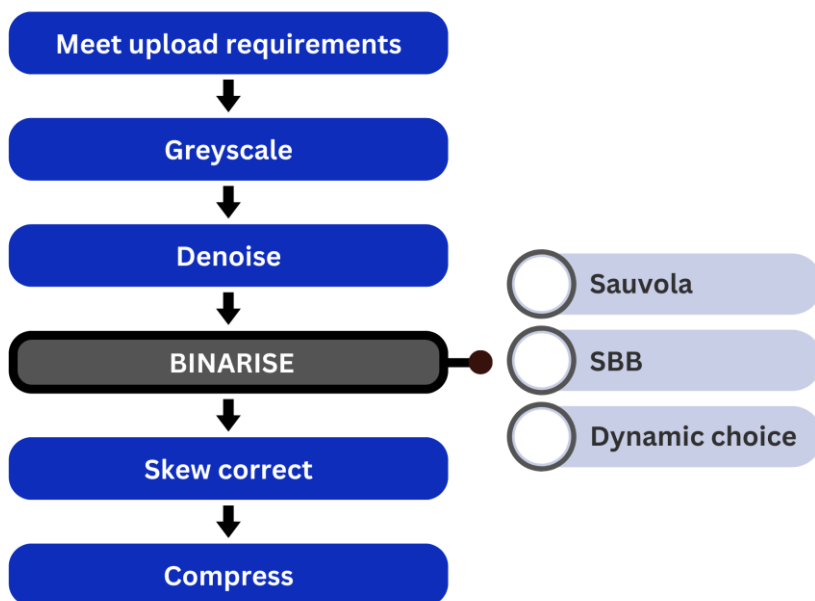


Figure 6 The foundational pipeline used for our three proposed methods. Each method uses a different binarisation method but otherwise shares the same pipeline.

For evaluation, we chose not to retrain our HTR model using the pre-processed images produced in our experiments. Instead, we evaluated pre-processing effects by inputting treated images into a model originally trained on untreated images. We hypothesised that this approach would enhance the model's robustness in dealing with image variation. In subsequent model iterations following our research, we incorporated both untreated and pre-processed images into our training set to enhance the model's robustness and ability to effectively generalise its learnings to unseen images by exposing the model to further variation.

### 3.1.3 *Method one: Sauvola binarisation pipeline*

The first approach in our pipeline (Fig. 6) employed Sauvola binarisation, a locally adaptive thresholding method particularly effective for addressing uneven illumination in historical images (Fig. 4, bottom). Unlike deep learning methods, which rely on diverse and representative training datasets, Sauvola binarisation applies fixed mathematical rules, producing consistent and predictable results regardless of the input image characteristics. This consistency made it a reliable choice for our dataset, despite its limitations.

Two settings, or hyperparameters, control Sauvola binarisation: window size and k-value. These parameters significantly influence the quality of binarisation, often requiring trade-offs between image subsets (Fig. 7).

For example, a larger window size enhances character definition but increases speckling in degraded images, while a higher k-value reduces noise but compromises contrast, especially in images with red text.

We tested k-values between 0.034 and 0.24, and window sizes from 11 to 41, ultimately selecting 0.14 for k-value and 21 for window size. This configuration maximised HTR accuracy across the dataset while minimising adverse effects on poor-quality and red-text images.

Despite these optimisations, red-text images remained particularly challenging, as boosting their contrast often degraded the quality of other subsets. The code for this method is publicly available (Sabbagh *et al.* 2024a).<sup>9</sup>

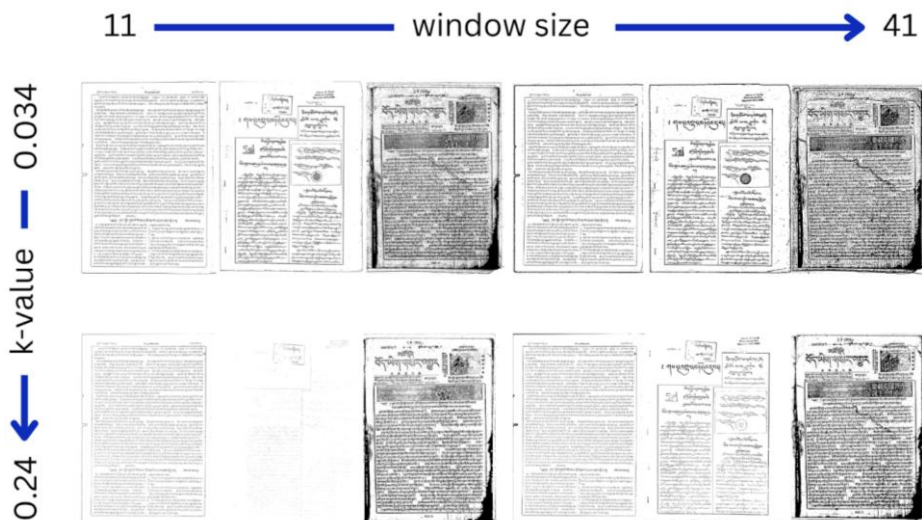


Figure 7 The trade-offs of using different Sauvola hyperparameter values (k-value, window size) in terms of image quality. The left-hand image of each trio is fair-quality, the middle image contains red text, and the right-hand image is poor-quality.<sup>1</sup> Each setting affects each image quality differently.

### 3.1.4 Method two: Deep learning binarisation pipeline (SBB binarisation)

The second method integrated the SBB deep learning-based binarisation model into our pipeline (Fig. 6). Unlike Sauvola binarisation, SBB employs a neural network to classify each pixel as foreground (text) or background by analysing patterns learned from training datasets, including those used in Document Image Binarization Contest (DIBCO) competitions, the Palm Leaf dataset

<sup>9</sup> Available online at [www.github.com/Divergent-Discourses/dd\\_preprocess](https://www.github.com/Divergent-Discourses/dd_preprocess) (accessed December 10, 2024).

(Burie *et al.* 2016), the Persian Heritage Image Binarization Competition (PHIBC) dataset (Ayatollahi & Nafchi 2013), and additional documents from the Berlin State Library (Rezanezhad 2023).<sup>10</sup>

The model analyses complex patterns, considering a pixel's immediate surroundings and its position within the broader context of the image when deciding whether it should be black or white. Since our Sauvola experiments (Section 3.1.3) indicated that tailoring binarisation settings by image type could improve results, SBB's adaptability offered a significant advantage. We anticipated that SBB would outperform traditional thresholding methods in handling complex challenges such as faded text, uneven illumination, and coloured or stained backgrounds.

### 3.1.5 *Method three: Forked binarisation pipeline*

Building on our evaluation of methods one and two, we developed a forked binarisation pipeline (Fig. 8) to combine their strengths and address the varying image qualities in our dataset. This approach aimed to improve HTR accuracy by dynamically selecting the most suitable binarisation method for each image. The code for this method is publicly available (Sabbagh *et al.* 2024b).<sup>11</sup>

Our analysis (Section 4) revealed distinct strengths for each method: Sauvola binarisation (method one) performed better on poor-quality images and those with red text, while SBB binarisation (method two) excelled with fair-quality images. However, some images achieved the highest HTR accuracy when left in their baseline, lightly processed state. Identifying these patterns motivated the creation of a pipeline capable of making automated decisions based on image quality. We therefore turned to image quality assessment (IQA) to identify

---

<sup>10</sup> More specifically, the model employs a hybrid CNN-Transformer architecture using a ResNet50-UNet encoder-decoder.

<sup>11</sup> Available at [www.github.com/Divergent-Discourses/dd\\_custom\\_preprocess](https://www.github.com/Divergent-Discourses/dd_custom_preprocess) (accessed December 10, 2024).

whether an image was fair-quality or poor-quality and pre-process this image using the pipeline best suited to its quality.

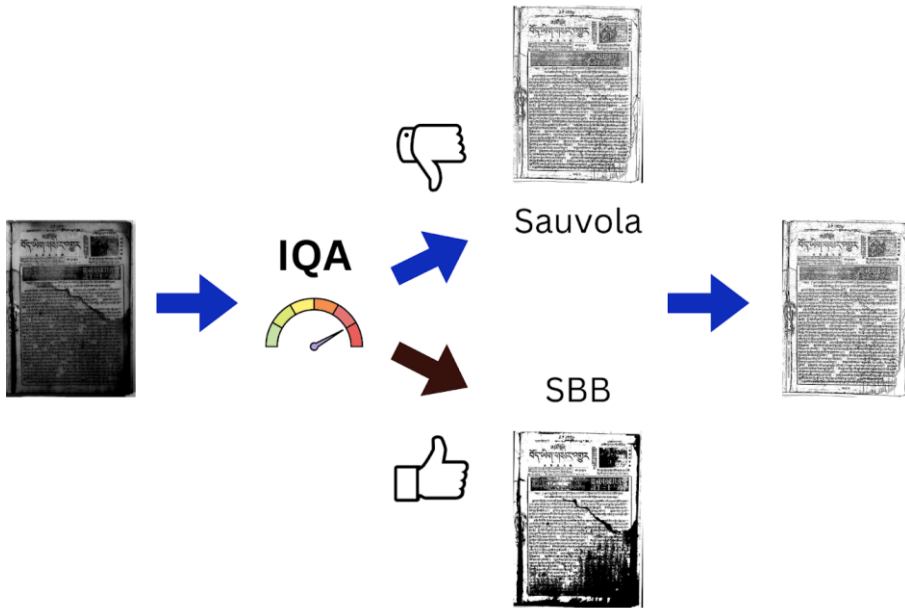


Figure 8 The forked binarisation pipeline and the potential path for one image<sup>1</sup> (blue arrows). An image quality assessment (IQA) method quality-scores an image. The image is considered poor-quality so the pipeline pre-processes it with the Sauvola binarisation-based pipeline. A score indicating good quality would have resulted in SBB binarisation-based pre-processing.

We used the MANIQA model (Yang *et al.* 2022), implemented via the PYIQA Python package (Chen & Mo 2021), to assign a perceptual quality score to each image, reflecting how a human might evaluate its visual quality. Images scoring above a threshold were classified as fair-quality and processed with SBB binarisation, while those scoring below the threshold were classified as poor-quality and processed with Sauvola binarisation. Appendix A details the rationale behind selecting MANIQA from among 18 tested IQA methods.

While this approach successfully classified images into two bins (fair- and poor-quality), it was unable to identify images that were better left in their baseline, lightly treated state. This issue stemmed from the binary nature of threshold-based classification, which

inherently divides images into only two classes: those scoring above or below the set threshold. Additionally, there were no consistently discernible visual characteristics or subjective criteria that reliably indicated when baseline, lightly treated images would outperform re-processed ones.

To mitigate this limitation, we decided to pre-process only a subset of the dataset—images from the library, Staatsbibliothek zu Berlin, comprising 9,377 of the 16,718 images in our total dataset. This subset predominantly consisted of poor-quality images<sup>12</sup> likely to benefit from pre-processing, along with a smaller number of good-quality images that appeared well-suited to SBB binarisation. It did not contain images with red text. Importantly, the subset appeared to contain few images that would have been most accurately transcribed in their baseline, lightly treated state, thereby minimising the risk of pre-processing inadvertently degrading transcription accuracy.

To adapt our approach to suit this library subset, we therefore adjusted Sauvola binarisation hyperparameters to better suit images without red text, setting the  $k$ -value to 0.24 and window size to 11. These settings improved binarisation quality for the remaining pages but were not ideal for the entire dataset, as discussed in Section 3.1.3. Using the forked pipeline, we applied Sauvola binarisation to poor-quality images and SBB binarisation to fair-quality images, predicting that this mixed approach would yield higher overall transcription accuracy.

Combining the forked pipeline evaluation with the new Sauvola hyperparameter settings presented a limitation: it was not possible to isolate the effects of the forked approach from those of the adjusted parameters. However, constraints within the Transkribus platform prevented us from conducting several separate evaluations. Despite this, the forked pipeline offers a pragmatic solution to the challenges posed by the varied quality of our dataset.

---

<sup>12</sup> These images are likely of poor quality due to the limitations of the original microfilm captures, produced by the China International Book Trading Corporation, Beijing, combined with the degradation of the microfilm itself over time.

### 3.2 *Evaluation methodology*

#### 3.2.1 *Test set composition*

To evaluate the methods described above, we curated a test set to assess HTR accuracy following pre-processing. The test set consisted of 86 images drawn from 11 newspapers, each manually transcribed to establish ground truth. This accounted for 0.5% of the total dataset. Although test sets typically represent a larger proportion of the overall data, expanding the test set was constrained by the time-intensive nature of manual transcription required to generate ground truth labels. A more standard 20% representation would have necessitated the transcription of approximately 3,400 images, which was not feasible within the scope of the project.

Instead, we prioritised diversity and ensured that the selected 86 images were representative of the key quality subsets identified within the dataset—specifically, poor-quality, fair-quality, and red-text images. Additionally, we targeted images exhibiting characteristics predicted to challenge HTR performance, such as bleed-through and underexposure. This approach allowed for a focused yet comprehensive evaluation of the model’s capabilities across the dataset’s most demanding scenarios.

To investigate how each method performed across the three main image quality categories in our dataset, we divided the test set into three subsets: fair-quality (42 images), poor-quality (29 images), and containing red text (15 images). These subsets were proportional to the estimated distribution of image qualities in the entire dataset, approximately 49%, 34%, and 17%, respectively. We did not use the test set images to train the HTR models.

Although we selected test set images to represent a wide range of characteristics hypothesised to contribute to HTR inaccuracies, the categorisation of images into the three subsets relied on subjective human judgment. These judgments may not have aligned perfectly with the features or patterns most relevant to deep learning algorithms. As a result, we may have placed some images in test subsets that did not align with how our HTR model interpreted them.

For example, an image the model struggled to transcribe may have been categorised as fair-quality based on human perception. This misalignment could lead to skewed subset results, potentially underestimating or overestimating the model's performance within specific categories. However, by ensuring that the selected images reflected diverse degradations, we aimed to mitigate this effect and capture the full spectrum of quality challenges.

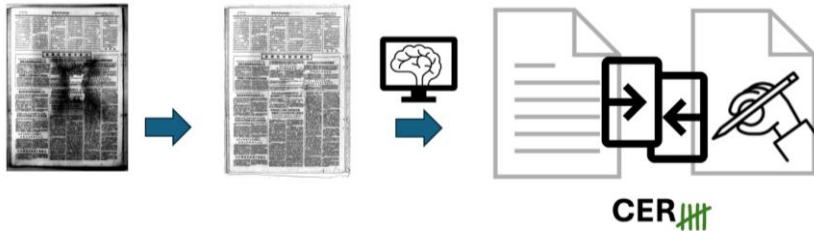
### 3.2.2 *Evaluating pre-processing methods*

To assess the performance of our three pre-processing methods, we executed the following (Fig. 9):

- (1) **Pre-processing:** For each method, we pre-processed the test set images, or for the baseline, ensured that the images met the minimum agreed requirements.
- (2) **HTR Processing:** The pre-processed images were inputted to the HTR model, which produced predicted transcriptions. Within Transkribus, we utilised two project-trained model prototypes: the field recognition model *TibNewsTR* and the handwritten text recognition (HTR) model *TibNewsOne4All 0.1*.
- (3) **CER Calculation:** For each image, we compared the predicted transcription with the ground truth to calculate the character error rate (CER) for each image (Appendix B). CER quantifies transcription errors - such as missing, incorrect, or additional characters - and ranges between 0 and 100, with higher scores indicating more errors. CER scores were computed using the Transkribus Expert client (Read-Coop SCE n.d.), as the standard Transkribus platform does not currently support this functionality (Transkribus n.d.).
- (4) **Averaging Results:** CER scores were averaged across subsets (fair-quality, poor-quality, red text) and across the entire test set to evaluate the effectiveness of each pre-processing method.



All CER scores were computed in a case-insensitive manner to avoid penalising transcription errors related to capitalisation, which did not make sense for Tibetan characters.



*Figure 9 Methodology to evaluate pre-processing methods. Page 4 of Qinghai Tibetan News, Jun. 14, 1955, is pre-processed with a given method (or left in its baseline, lightly processed state). The pre-processed image is inputted to the HTR model which outputs a predicted transcription. The predicted transcription is compared to the 'ground truth' transcription, outputting a character error rate (CER) value.*

During our evaluations, the Divergent Discourses project continued to develop new iterations of its HTR model. To ensure consistency, we selected a specific model iteration for all evaluations. Since then, the project has produced newer HTR model versions, trained on more diverse datasets and incorporating additional HTR steps (e.g., line polygon detection). As a result, our findings on the most effective pre-processing methods may not directly apply to the latest HTR models. This highlights a limitation of conducting evaluations alongside the ongoing development of interconnected components such as HTR models.

#### 4 Results

The experiments aimed to evaluate whether image pre-processing improved HTR transcription accuracy on our dataset of historical Tibetan newspaper pages. Additionally, we sought to determine which pre-processing method achieved the most accurate transcriptions overall, while considering that different image qualities might require distinct treatments.

This section first presents the visual outcomes of each pre-

processing method compared to the baseline. Next, it discusses transcription error rates for the entire test set, and finally examines transcription error rates for the subset of data selected for pre-processing based on the results of the first two pipelines.



Figure 10 Visual outcomes and character error rate (CER) values when image is left in baseline, lightly treated state and after pre-processing using Sauvola and SBB binarisation-based pipelines. One image is shown for each test subset: fair-quality, poor-quality, and containing red text.<sup>13</sup>

<sup>13</sup> Defend Tibet's Freedom, Aug. 20, 1963, p.4 (fair-quality); Qinghai Tibetan News, Jul. 5, 1952, p.3 (poor-quality); News in Brief, Dec. 1, 1953, p.4 (red text).

#### 4.1 *Image pre-processing outcomes*

Visual outcomes of each pre-processing method, alongside baseline, lightly treated images, are shown in Figure 10. These examples illustrate how each method transformed the images prior to HTR processing.

#### 4.2 *Transcription error rates across test set*

Table 1 presents transcription error rates, measured using character error rate (CER), for each pre-processing method compared to the baseline. Median CER values suggest that all pre-processing methods improved HTR accuracy relative to the baseline. This trend was consistent with mean CER values, except for fair-quality images, where the mean for baseline images (38.89) slightly outperformed pre-processed results (39.86).

SBB binarisation yielded the lowest **median** CER for fair-quality images, suggesting that SBB binarisation was the most effective pre-processing method for fair-quality images. Fair-quality images left in their baseline, lightly pre-processed state had the lowest **mean** CER, suggesting that it was most effective to leave these images in their baseline, lightly treated state. While these findings are contradictory, median CER values and the resulting conclusions are likely to be more representative of performance, given the presence of outliers that skewed the data. This issue is explored further in Section 5.2. Sauvola binarisation was the most effective for poor-quality images and those containing red text, achieving the lowest CER values according to both mean and median values.

Across the entire test set, the forked binarisation method resulted in the lowest CER values, as hypothesised. Mean (39.91) and median (40.08) CER values for the forked method were both lower than those for any other method or the baseline, which ranged from 40.24 to 43.47 (mean) and 41.19 to 42.47 (median).

In general, transcription accuracy varied by image type, with red text images achieving the lowest CER scores (indicating the best

accuracy) across all methods, followed by fair-quality images, and finally poor-quality images.

*Table 1 Character error rates (CER) across test subsets (fair-quality, poor-quality, red text) and our entire test set. The best-performing method for each subset is shown in bold*

Pre-processing Method	Fair-quality (CER) ↓		Poor-quality (CER) ↓		Red text (CER) ↓		Overall (CER) ↓	
	Median	Mean	Median	Mean	Median	Mean	Median	Mean
Baseline	41.62	<b>38.89</b>	56.85	55.13	33.14	28.01	43.47	42.47
Method 1 (Sauvola binarisation)	43.06	44.42	<b>45.48</b>	<b>44.99</b>	<b>19.79</b>	<b>25.15</b>	40.24	41.25
Method 2 (SBB binarisation)	<b>39.86</b>	39.32	51.82	50.08	29.36	29.24	41.85	41.19
Method 3 (forked binarisation)	43.12	39.56	50.54	46.00	29.33	29.13	<b>40.08</b>	<b>39.91</b>

These results are illustrated in Figure 11, which shows the distribution of image-wise CER scores for each method. Baseline, lightly treated images generally had higher error rates – particularly for poor-quality images – with a wide spread of data points, reflecting the challenges faced by the HTR model in reliably identifying text in non-pre-processed images.

Both Sauvola binarisation and SBB binarisation achieved lower mean error rates compared to baseline, lightly treated images. Sauvola demonstrated greater improvement for poor-quality images, while SBB slightly outperformed for fair-quality images. Sauvola appeared to reduce the accuracy of some fair-quality images relative to leaving the images in their baseline, lightly processed state.

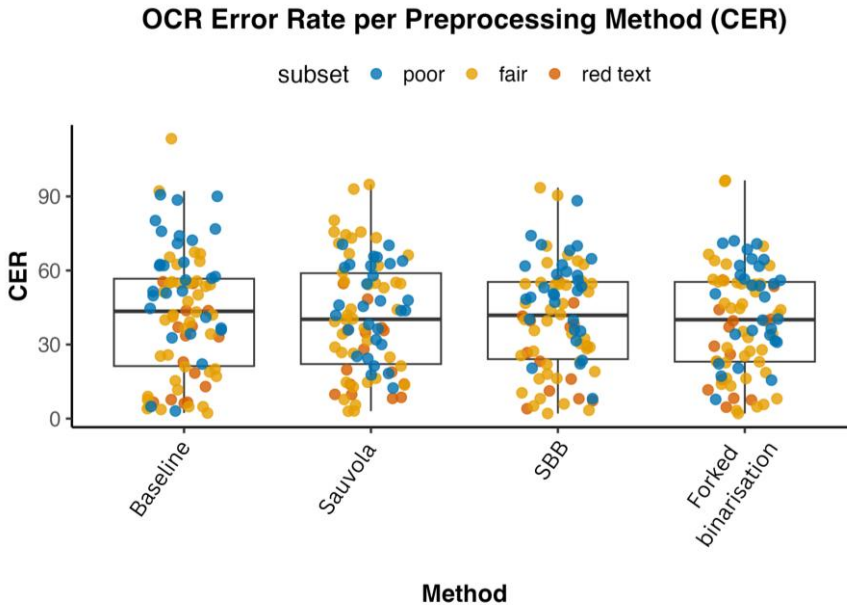


Figure 11 A box and whisker plot illustrating the distribution of individual image character error rates (CER) across methods. Values are shown for poor-quality (blue), fair-quality (orange) and red text (red) images.

The forked binarisation method showed the best overall performance, with the lowest mean error rate and tighter clustering of data points, indicating more consistent results across the test set. Furthermore, it produced fewer outliers with high error rates, reinforcing its effectiveness in preparing diverse image qualities for HTR.

### 4.3 Transcription error rates across library images

The test set included 27 images from the library, Staatsbibliothek zu Berlin, whose dataset we selected for pre-processing. Of these, 4 were fair-quality (15%), 23 were poor-quality (85%), and none contained red text. A qualitative review of the full dataset from this library suggests it contains a higher proportion of fair-quality images than represented in this test subset. Thus, while not strictly representative, this subset was evaluated to investigate whether the forked binarisation pipeline

would result in more accurate transcriptions than single-method approaches, as hypothesised.

Table 2 *Character error rates (CER) across images from the library, Staatsbibliothek zu Berlin, selected for pre-processing. Values are shown for test subsets (fair-quality, poor-quality) and overall. The best-performing method for each subset is highlighted in bold.*

Pre-processing Method	Fair-quality (CER) ↓		Poor-quality (CER) ↓		Overall (CER) ↓	
	Median	Mean	Median	Mean	Median	Mean
Baseline	41.62	43.66	61.90	61.87	56.85	59.17
Method 1 (Sauvola binarisation)	<b>20.56</b>	<b>26.01</b>	<b>47.70</b>	<b>48.03</b>	<b>45.98</b>	<b>44.77</b>
Method 2 (SBB binarisation)	22.67	28.89	53.50	55.01	53.04	51.14
Method 3 (forked binarisation)	21.77	27.74	53.95	49.51	53.79	46.28

Table 2 presents the results. Sauvola binarisation achieved the most accurate transcriptions for fair-quality, poor-quality, and the overall subset in terms of both mean and median CER values. All pre-processing methods improved transcription accuracy relative to the baseline, lightly treated images.

While the forked method (Method 3) was expected to outperform single-method approaches, the results showed Sauvola binarisation consistently achieved the lowest CER for all subsets.

## 5 Discussion

The results demonstrated that image pre-processing generally improved transcription accuracy compared to using baseline, lightly

treated images, with some exceptions for fair-quality images (Table 1, Table 2). However, the findings also highlighted that different image qualities benefited from distinct pre-processing treatments. This underscores the potential advantage of adaptive or multi-path pipelines that dynamically tailor pre-processing approaches based on detected image attributes. Notably, the forked binarisation pipeline delivered the highest overall transcription accuracy for a dataset containing images with diverse characteristics, validating its adaptive approach.

The SBB binarisation pipeline excelled in transcribing fair-quality images but struggled with poor-quality ones, performing worse than Sauvola binarisation. SBB’s deep learning approach appeared to adapt well to the specific features of fair-quality images, producing cleaner backgrounds and more defined foreground text (Fig. 12).

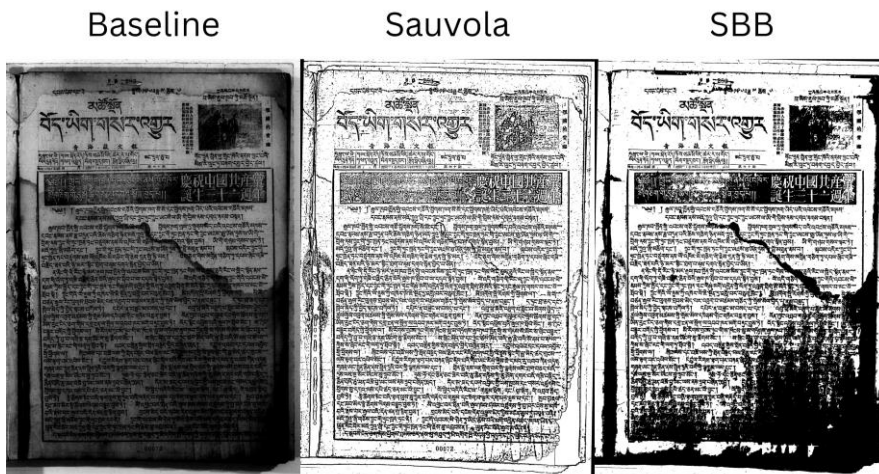


Figure 12 Visual effects of Sauvola and SBB binarisation compared to the baseline, lightly treated image. SBB binarisation resulted in cleaner, less speckled backgrounds but handled staining and dark patches poorly resulting in information loss.

However, it handled staining and dark patches poorly, often rendering these areas entirely black and causing significant information loss. This limitation is likely attributable to the training data for the SBB model, which may have predominantly consisted of images similar in quality to those classified as fair in our dataset. The model’s lack of

exposure to poor-quality images during training likely limited its ability to address issues such as staining.

Sauvola binarisation, by contrast, was more effective for poor-quality images and those containing red text. Sauvola generally preserved more information in stained and distorted areas, minimising the risk of rendering significant portions of the page unusable. However, in some cases, SBB binarisation unexpectedly outperformed Sauvola for poor-quality images. These cases often involved minimal staining and significant bleed-through, where SBB's handling of the bleed-through yielded better results (Fig. 13).



Figure 13 *Sauvola binarisation used hyperparameters ( $k$ -value and window size) optimised for all three test subsets (fair-quality, poor-quality, red text) - the trade-off resulted in poorly binarised bleed-through relative to SBB binarisation (Qinghai Tibetan News, Jul. 12 1952 p2).*

Conversely, when low-resolution, blotchy characters were present, Sauvola occasionally outperformed SBB even for mostly fair-quality images. In such instances, SBB's tendency to merge stacked characters and increase blotchiness reduced transcription accuracy despite its ability to produce less speckled backgrounds (Fig. 14).





Figure 14 SBB binarisation sometimes resulted in character blotchiness rendering characters less legible than their Sauvola-binarised counterparts, even in images considered to be fair-quality (Tibet Daily, Jan. 1 1962, p1).

For red text, Sauvola binarisation generally resulted in superior transcription accuracy. This was despite SBB producing higher-contrast text, whereby achieving similar contrast with Sauvola would require fixed settings that would risk degrading pre-processing quality for other image types. Additionally, SBB often degraded page details, such as column borders, that may have served as useful clues for text region detection within the HTR pipeline. Sauvola preserved these details more consistently, contributing to its superior performance in these cases.

Interestingly, the baseline (lightly treated) inputs outperformed pre-processing for a subset of images, primarily from the fair-quality and red-text categories. Among these, 55% were fair-quality, a disproportionately high representation compared to the overall test set, while 24% were red text, also higher than expected. In contrast, only 21% were poor-quality, a lower proportion than their overall representation. These findings suggest that pre-processing is not universally beneficial, and its impact depends heavily on the specific characteristics of the input data.

For poor-quality images, pre-processing was particularly advantageous, likely due to its ability to correct distortions and enhance features which assisted HTR. For fair-quality images or those with specific challenges, such as red text, baseline (lightly treated) inputs sometimes retained details that pre-processing inadvertently degraded. For example, images with colour content risked reduced contrast when applying treatments optimised for other image types. This variability highlights the importance of selective or adaptable pre-processing approaches. Tailoring pre-processing to specific image attributes—such as quality, presence of red text, or colour—may mitigate issues of information loss and further optimise HTR performance.

### 5.1 *Forked binarisation method*

Recognising the benefits of tailoring pre-processing to specific image characteristics, our forked binarisation method dynamically selected the most suitable pre-processing approach to optimise transcription accuracy. This strategy aligns with the multi-pass approach described by Chastagnol (2013), who addressed the challenges of designing pipelines for heterogeneous datasets in a commercial setting.

Chastagnol's multi-pass algorithm applied several pre-processing methods to each image, quality-scored the results, and selected the highest-scoring version for HTR. In contrast, our approach is more computationally efficient: by quality-scoring images upfront, we determine the optimal pre-processing method, reducing the number of operations while still maintaining a focus on enhancing transcription accuracy.

On the overall test set, our approach's adaptability to image characteristics resulted in higher overall transcription accuracy compared to using a single pre-processing method. The image-wise evaluation data indicated that optimising hyperparameters for Sauvola binarisation further improved accuracy for pages where Sauvola was the preferred method. For images where the pipeline correctly selected Sauvola, the character error rate (CER) values were

generally lower than those achieved with our formerly-selected Sauvola hyperparameter values, except for on images containing red text. For red-text images, CER values were higher than expected due to the hyperparameters having been optimised for non red-text images. If the pipeline was enhanced to also leave images in their baseline (lightly treated) state when beneficial, transcription accuracy would likely improve further.

On the Staatsbibliothek zu Berlin image test set, the adaptability of the forked binarisation pipeline did not yield higher overall transcription accuracy. This is likely because 85% of the images were poor-quality, which would predominantly benefit from Sauvola binarisation based on our results. Given the relatively homogeneous nature of this test set, the dynamic approach offered by the forked pipeline was not necessary and generated errors. However, the library test set did not fully represent the broader image quality distribution from the library. A qualitative review suggests that the actual library image set includes a higher proportion of fair-quality images, which would likely benefit more from the dynamic selection provided by the forked binarisation pipeline.

A limitation of our approach was that the choice between Sauvola and SBB binarisation did not strictly correlate with the image quality categories (e.g., fair or poor). These quality labels served as proxies for the true objective: determining the optimal pre-processing method (either Sauvola or SBB) for a given image. The reasons why some images are more accurately transcribed with one method over the other remain speculative and are likely not strictly related to image quality. Ideally, a neural network could be trained to predict the most suitable pre-processing method for each image, effectively emulating the decision-making process of our approach. However, generating sufficient training data for this task was outside the scope of this project.

### 5.2 *Data outliers*

As discussed in Section 4.2, our dataset contained a small number of outliers that made the median CER values more reliable than the mean. These outliers included images that were automatically transcribed significantly more accurately in their baseline, lightly treated form than after applying any of the three pre-processing methods. We suspect these images were inadvertently included both in the training set for our HTR model and in the test set. As a result, the model likely performed better on the baseline (lightly treated) images because it had been trained on these exact baseline images and their corresponding transcriptions, essentially "memorising the answers" during training. In contrast, the pre-processed versions of these images, which the model had not encountered during training, were transcribed with greater error, highlighting the importance of proper dataset partitioning in machine learning model development.

Another source of outliers was pages containing a significant amount of Chinese text. Our HTR model had been trained primarily on Tibetan text with limited exposure to Chinese. Consequently, the transcription accuracy for these pages was lower than for others in the dataset. Despite these limitations, these pages containing Chinese language still enabled valid comparisons across pre-processing methods, as they were consistently included in all experiments.

### 5.3 *Future work*

Future work could explore more advanced approaches to tailoring pre-processing methods within the pipeline while remaining mindful of budgetary constraints typical of project-based work. One promising avenue is the use of clustering algorithms to automatically group images by quality, enabling the application of targeted pre-processing strategies to each cluster. Additionally, computer vision techniques could be employed to identify specific problem areas within images—such as staining, uneven illumination, or coloured text—and customise pipelines for individual images.

For example, the Turing Institute’s MapReader tool (Wood *et al.* 2024) classifies patches of an image and could potentially be adapted to identify visual features in historical newspaper images. This approach aligns with budgetary constraints, as it would require relatively simple patch-wise annotation of visual features rather than the resource-intensive task of manual transcription for training a model.

It is also important to recognise that achieving a CER of zero is rare; HTR transcriptions will almost always contain some degree of imperfection. Researchers must account for such errors—or ‘HTR noise’—when using these transcriptions in downstream applications, such as building databases or conducting computational textual analysis. For instance, the Impresso project addresses optical character recognition (OCR) noise by offering a keyword suggestion tool that proposes fuzzy matches for user search queries, ensuring that relevant texts are retrieved even when transcription errors occur (Düring *et al.* 2024). Similar tools could be developed to support Tibetan studies, where HTR challenges are compounded by the complexity of the script and the variability of historical document conditions.

## 6 Conclusion

Our research addressed the challenge of enhancing HTR transcription accuracy for historical Tibetan newspaper images through tailored and adaptive pre-processing strategies. By evaluating three distinct binarisation approaches, we demonstrated the important role of context- and quality-aware image enhancement techniques in improving HTR outcomes for heterogeneous document collections.

The forked binarisation method we developed offers a promising solution to the complexities of digitising heterogeneous historical texts. Unlike uniform pre-processing strategies, our approach dynamically selects the most appropriate method based on individual image characteristics, striking a balance between accuracy and computational efficiency.

Beyond the technical advancements, our work underscores the broader implications of HTR quality. Variations in document condition can introduce biases in the accessibility of historical texts, potentially affecting socio-political research and cultural preservation efforts. This highlights the importance of nuanced, adaptive digitisation approaches that consider the importance of technical performance in relation to evenly distributed access.

By openly sharing our method, we aim to support future research in Tibetan studies, document preservation, and the broader field of historical HTR. Refining these adaptive pre-processing strategies could significantly enhance access to historical texts that might otherwise remain inaccurately transcribed and overlooked, thereby supporting their continued study and preserving their cultural significance.

### Bibliography

- Alaei, Alireza, Vinh Bui, David Doermann, and Umapada Pal  
"Document Image Quality Assessment: A Survey," *ACM Computing Surveys*, 56 (2), 2023, pp. 1–36. [doi: 10.1145/3606692](https://doi.org/10.1145/3606692)
- Anvari, Zahra, and Vassilis Athitsos  
"A Survey on Deep Learning Based Document Image Enhancement," *arXiv preprint*, 2021. [doi:10.48550/arXiv.2112.02719](https://doi.org/10.48550/arXiv.2112.02719)
- Ayatollahi, Seyed Morteza, and Hossein Ziaei Nafchi  
"Persian heritage image binarization competition (PHIBC 2012)." In *First Iranian Conference on Pattern Recognition and Image Analysis (PRIA)*, IEEE, 2013, pp. 1-4. [doi:10.1109/PRIA.2013.6528442](https://doi.org/10.1109/PRIA.2013.6528442)
- Beelen, Kaspar, Jon Lawrence, Daniel C.S. Wilson, and David Beavan  
"Bias and Representativeness in Digitized Newspaper Collections: Introducing the Environmental Scan," *Digital Scholarship in the Humanities* 38 (1), 2023, pp. 1–22. [doi:10.1093/llc/fqac037](https://doi.org/10.1093/llc/fqac037)

Bhattacharai, Ashuta (username ashuta03)

"Automatic Skew Correction Using Corner Detectors and Homography," *GitHub repository*, 2019. Available online at <https://github.com/ashuta03/automatic-skew-correction-using-corner-detectors-and-homography> (accessed July 3, 2024).

Bradski, Gary

"The OpenCV Library," *Dr. Dobb's Journal of Software Tools* 25 (11), 2000, pp. 120, 122-125.

Brunner, Stéphane

"Deskew: Skew detection and correction in images containing text," *Python Package Index (PyPI)*, 2024. Available online at <https://pypi.org/project/deskew/> (accessed July 3, 2024).

Burie, Jean-Christophe, Mickaël Coustaty, Setiawan Hadi, *et al.*

"ICFHR2016 competition on the analysis of handwritten text in images of Balinese palm leaf manuscripts." In *15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, IEEE, 2016, pp. 596-601. [doi:10.1109/ICFHR/2016.107](https://doi.org/10.1109/ICFHR/2016.107)

Chastagnol, François

"Building an Image Pre-processing Pipeline in Python," *YouTube video, Next Day Video*, 2013. Available online at <https://www.youtube.com/watch?v=B1d9dpqBDVA> (accessed October 29, 2024).

Chen, Chaofeng, and Jiadi Mo

"IOA-PyTorch: PyTorch Toolbox for Image Quality Assessment," 2021. Available online at <https://github.com/chaofengc/IOA-PyTorch> (accessed October 29, 2024).

Coutts, Margaret

*Stepping Away from the Silos: Strategic Collaboration in Digitisation*. Chandos Publishing, 2016.

## DIBCO

"Datasets," Last updated October 4, 2023. Available online at <https://dib.cin.ufpe.br/#!/datasets> (accessed October 29, 2024).

Sabbagh, Christina, Franz Xaver Erhard, Robert Barnett, and Nahan W. Hill

"Divergent Discourses Custom Image Preprocessing (Sauvola Binarisation)," *Zenodo*, 2024a. [doi:10.5281/zenodo.14525692](https://doi.org/10.5281/zenodo.14525692).

"Divergent Discourses Custom Image Preprocessing (Forked Binarisation)," *Zenodo*, 2024b. [doi:10.5281/zenodo.14523007](https://doi.org/10.5281/zenodo.14523007).

Düring, Marten, Estelle Bunout, and Daniele Guido

"Transparent Generosity: Introducing the impresso Interface for the Exploration of Semantically Enriched Historical Newspapers," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 2024, pp. 35–55. [doi:10.1080/01615440.2024.2344004](https://doi.org/10.1080/01615440.2024.2344004)

Erhard, Franz Xaver

"The Divergent Discourses Corpus: A Digital Collection of Early Tibetan Newspapers of the 1950s and 1960s," *Revue d'Études Tibétaines*, (74), 2025, pp. 45–81.

Ehrmann, Maud, Edouard Bunout, and Frédéric Clavert

"Digitised Historical Newspapers: A Changing Research Landscape." In *Newspapers—A New Eldorado for Historians*, 2023, pp. 1–22. [doi:10.1515/9783110729214-001](https://doi.org/10.1515/9783110729214-001)

Griffiths, Rachael

"Handwritten Text Recognition (HTR) for Tibetan Manuscripts in Cursive Script," *Revue d'Études Tibétaines*, (72), 2024, pp. 43–51. [doi:10.1553/TibSchol\\_ERC\\_HTR](https://doi.org/10.1553/TibSchol_ERC_HTR).

Gupta, Maya, R, Nathaniel P. Jacobson, and Erik K. Garcia

"OCR Binarization and Image Pre-Processing for Searching Historical Documents," *Pattern Recognition* 40 (2), 2007, pp. 389–397. [doi:10.1016/j.patcog.2006.04.043](https://doi.org/10.1016/j.patcog.2006.04.043).



- Hosu, Vlad, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe  
"KonIQ-10k: An Ecologically Valid Database for Deep Learning of Blind Image Quality Assessment," *IEEE Transactions on Image Processing* 29, 2020, pp. 4041–4056. [doi:10.1109/TIP.2020.2967829](https://doi.org/10.1109/TIP.2020.2967829)
- Jacsont, Pauline, and Elina Leblanc  
"Impact of Image Enhancement Methods on Automatic Transcription Trainings with eScriptorium," *Journal of Data Mining & Digital Humanities*, 2023. [doi:10.46298/jdmdh.10262](https://doi.org/10.46298/jdmdh.10262)
- Lins, Rafael Dueire, Rodrigo B. Bernardino, Edward B. Smith, *et al.*  
"ICDAR 2021 Competition on Time-Quality Document Image Binarization." In *Document Analysis and Recognition – ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part IV, vol. 16*, Springer International Publishing, 2021, pp. 708–22. [doi:10.1007/978-3-030-86337-1\\_47](https://doi.org/10.1007/978-3-030-86337-1_47)
- Luo, Queenie, and Leonard W.J. van der Kuip  
"Norbu Ketaka: Auto-Correcting BDRC's E-Text Corpora Using Natural Language Processing and Computer Vision Methods," *Revue d'Études Tibétaines* (72), 2024, pp. 26–42. Available online at [https://d1i1jdw69xsqx0.cloudfront.net/digitalhimalaya/collections/journals/ret/pdf/ret\\_72\\_02.pdf](https://d1i1jdw69xsqx0.cloudfront.net/digitalhimalaya/collections/journals/ret/pdf/ret_72_02.pdf) (accessed January 26, 25).
- Niblack, Wayne  
*An Introduction to Digital Image Processing*. Englewood Cliffs: Prentice-Hall, 1986.
- Nockels, Joseph, Paul Gooding, and Melissa Terras  
"The Implications of Handwritten Text Recognition for Accessing the Past at Scale," *Journal of Documentation* 80 (7), 2024, pp. 148–167. [doi:10.1108/JD-09-2023-0183](https://doi.org/10.1108/JD-09-2023-0183)
- Otsu, Nobuyuki  
"A Threshold Selection Method from Gray-Level Histograms," *Automatica* (11), 1975, pp. 23–27.

Panzer, Jason Drew

"Hough Transform implementation," *GitHub*, 2017. Available online at <https://gist.github.com/panzerama/beebb12a1f9f61e1a7aa8233791bc253> (accessed July 3, 2024).

Rawat, Sukhbindra Singh; Ashutosh Sharma, and Rachana Gusain

"Analysis of Image Pre-processing Techniques to Improve OCR of Garhwali Text Obtained Using the Hindi Tesseract Model," *ICTACT Journal on Image & Video Processing*, 12 (2), 2021. [doi:10.21917/ijivp.2021.0366](https://doi.org/10.21917/ijivp.2021.0366)

Read-Coop SCE

"Transkribus Expert Client, Version 1.28.0," Software. n.d. Available online at <https://readcoop.eu/transkribus/> (accessed November 3, 2024).

Reddy, Susmith

"Pre-Processing in OCR!!!" *Towards Data Science*, 2019. Available online at <https://towardsdatascience.com/pre-processing-in-ocr-fc231c6035a7>. (accessed July 3, 2024).

Rezanezhad, Vahid, Konstantin Baierer, and Clemens Neudecker

"A Hybrid CNN-Transformer Model for Historical Document Image Binarization," In Antonacopoulos, Apostolos, Christian Clausner, Maud Ehrmann, Kai Labusch, and Clemens Neudecker (eds.) *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing (HIP) 2023*, San José, 2023. [doi: 10.1145/3604951.3605508](https://doi.org/10.1145/3604951.3605508).

Sauvola, Jaakko, J. and Matti K. Pietikainen

"Adaptive Document Image Binarization," *Pattern Recognition* 33, 2000, pp. 225–236, [doi:10.1016/S0031-3203\(99\)00055-2](https://doi.org/10.1016/S0031-3203(99)00055-2).

Smith, David A., and Ryan Cordell

"A Research Agenda for Historical and Multilingual Optical Character Recognition," *NULab*, Northeastern University, 2018, pp. 36. [doi:10.1177/0961000611434760](https://doi.org/10.1177/0961000611434760).

Smith, Lucy, and Jennifer Rowley

"Digitisation of Local Heritage: Local Studies Collections and Digitisation in Public Libraries," *Journal of Librarianship and Information Science*, 44 (4), 2012, pp. 272–80.

Taş, İdal Çetin and Ahmet Anil Müngen

"Using Pre-Processing Methods to Improve OCR Performances of Digital Historical Documents." In *Innovations in Intelligent Systems and Applications Conference (ASYU)*, IEEE, 2021, pp. 1–5. [doi:10.1109/ASYU52992.2021.9598972](https://doi.org/10.1109/ASYU52992.2021.9598972).

Transkribus

"6. Computing Accuracy", n.d. Available online at <https://help.transkribus.org/computing-accuracy> (accessed December 7, 2024).

Wood, Rosie, Kasra Hosseini, Kalle Westerling, *et al.*

"MapReader: Open Software for the Visual Analysis of Maps," *Journal of Open Source Software* 9 (101), 2024, p. 6434. [doi:10.21105/joss.06434](https://doi.org/10.21105/joss.06434).

Yang, Sidi; Tianhe Wu, Shuwei Shi, Shanshan Lao, *et al.*

"Maniqa: Multi-Dimension Attention Network for No-Reference Image Quality Assessment." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1191–1200. [doi:10.1109/CVPRW56347.2022.00126](https://doi.org/10.1109/CVPRW56347.2022.00126)

Zhou, Yanxi, Shikai Zuo, Zhengxian Yang, Jinlong He, Jianwen Shi, and Rui Zhang

"A Review of Document Image Enhancement Based on Document Degradation Problem," *Applied Sciences*, 13 (13), 2023, p. 7855. [doi:10.3390/app13137855](https://doi.org/10.3390/app13137855).

## Appendices

### *Appendix A Selecting an image quality assessment (IQA) method*

To select the most effective Image Quality Assessment (IQA) method for use in our forked binarisation pipeline (Section 3.1.5), we conducted an experiment using the IQA methods available through the PYIQA Python package (Chen *et al.* 2021). Our goal was to determine which method would most accurately assess the quality of images in our test set.

We began by evaluating the performance of three methods; the first leaving the images in their baseline (lightly treated) state, the second applying the Sauvola binarisation method (Section 3.1.3), and the third applying the SBB binarisation method (Section 3.1.4). These evaluations provided image-wise Character Error Rates (CER) for each image, with three CER values per image - one for each method - indicating how accurately our HTR model transcribed each version of the image. Using these CER values, we classified each image according to the method that resulted in the most accurate transcription.

Since the IQA methods output a quality score for each image, we categorised the images into two classes based on a chosen threshold: 'fair-quality' and 'poor-quality'. Images most accurately transcribed by the Sauvola binarisation pipeline were labelled as 'poor-quality', while those best transcribed by the SBB binarisation pipeline were labelled as 'fair-quality'. For images which were most accurately transcribed when left in their baseline (lightly treated) state, we labelled them according to the second-best method.

Using each IQA method (listed in Table 3), we obtained quality scores for the images in our test set. We then calculated the mean quality score across the dataset and used this as the threshold to predict whether each image should be labelled as 'fair-quality' or 'poor-quality'. This process was repeated for all 18 methods, and we tracked the number of correct classifications across the entire dataset, as well as within each quality subset (fair-quality, poor-quality, red

text). This ensured that the methods did not perform disproportionately well on one subset and poorly on others.

Table 3 IQA model variants included in our experiments, alongside results over our entire test set. Names following dashes refer to the datasets upon which models were trained on.

Method Number	Method Name	Correctly predicted labels (/86)	Model Source
1	ARNIQA-clive	45	Agnolucci <i>et al.</i> 2024
2	ARNIQA-csiq	37	
3	ARNIQA-flive	39	
4	ARNIQA-kadid	50	
5	ARNIQA-koniq	42	
6	ARNIQA-live	42	
7	ARNIQA-spaq	42	
8	ARNIQA-tid	47	
9	BRISQUE	42	Mittal <i>et al.</i> 2012
10	CNNIQA	51	Kang <i>et al.</i> 2014
11	DBCNN	47	Zhang <i>et al.</i> 2018
12	HyperIQA	50	Su <i>et al.</i> , 2020
13	MANIQA-kadid	51	Yang <i>et al.</i> 2022
14	MANIQA-koniq	<b>53</b>	
15	MANIQA-pipal	52	
16	TReS-flive	33	Golestaneh <i>et al.</i> 2022
17	TReS -koniq	50	
18	WaDIQaM-nr	33	Bosse <i>et al.</i> 2017

Among the IQA methods tested, the quality scores from MANIQA-koniq resulted in the highest number of correctly classified images. As a result, we incorporated this model into our forked binarisation pipeline and adjusted the threshold from 0.2 to 0.335, which appeared to be more suitable for classifying our particular dataset.

*Appendix B Character error rate (CER)*

In this study, character error rate (CER) represents the percentage of characters which the HTR model has transcribed incorrectly, as determined through comparison with a 'ground truth', or reference, transcription. A lower CER value indicates a higher accuracy HTR model. It can be calculated as shown below.

N is the number of characters in the ground truth transcription. S is the number of character substitutions, D is the number of character deletions and I is the number of character insertions relative to the 'ground truth'.

$$CER = \frac{S + D + I}{N} \times 100$$

A character would be considered a substitution when incorrect but corresponding to one character in the ground truth (e.g. if the HTR model predicts the word ལྷལ but the correct transcription is ལྷལ།). A deletion would be where a character was missing from the predicted transcription (e.g. ལྷ). An insertion would be where the HTR model incorrectly predicted an additional character.

**Supplementary bibliography**

Agnolucci, Lorenzo, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo

"Arniqa: Learning Distortion Manifold for Image Quality Assessment." In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 189–198 [doi:10.1109/WACV57701.2024.00026](https://doi.org/10.1109/WACV57701.2024.00026)

Bosse, Sebastian, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek

"Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment," *IEEE Transactions on Image Processing* 27 (1), 2017, pp. 206–219. [doi:10.1109/TIP.2017.2760518](https://doi.org/10.1109/TIP.2017.2760518)

- Golestaneh, S. Alireza, Saba Dadsetan, and Kris M. Kitani  
“No-Reference Image Quality Assessment via Transformers, Relative Ranking, and Self-Consistency.” In *IEEE Computer Society*, 2022, pp. 3989–3999. [doi:10.1109/WACV51458.2022.00404](https://doi.org/10.1109/WACV51458.2022.00404)
- Kang, Le, Peng Ye, Yi Li, and David Doermann  
“Convolutional Neural Networks for No-Reference Image Quality Assessment.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1733–1740. [doi:10.1109/CVPR.2014.224](https://doi.org/10.1109/CVPR.2014.224)
- Mittal, Anish., Anush Krishna Moorthy, and Alan Conrad Bovik  
“No-reference image quality assessment in the spatial domain,” *IEEE Transactions on image processing*, 21 (12), 2012, pp. 4695–4708. [doi:10.1109/TIP.2012.2214050](https://doi.org/10.1109/TIP.2012.2214050).
- Su, Shaolin, Qingsen Yan, Yu Zhu, *et al.*  
“Blindly Assess Image Quality in the Wild Guided by a Self-Adaptive Hyper Network.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3667–3676. [doi:10.1109/CVPR42600.2020.00372](https://doi.org/10.1109/CVPR42600.2020.00372)
- Zhang, Weixia, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang  
“Blind Image Quality Assessment Using a Deep Bilinear Convolutional Neural Network,” *IEEE Transactions on Circuits and Systems for Video Technology* 30 (1), 2018, pp. 36–47. [doi:10.1109/TCSVT.2018.2886771](https://doi.org/10.1109/TCSVT.2018.2886771)

