Innovative Binarization Solutions for Historical Document Clarity

Radhika V. Kulkarni   
*Department of Computer Engineering*  
*Vishwakarma Institute of Technology*Pune, India  
radhika.kulkarni@vit.edu

Aarya Nirgude   
*Department of Computer Engineering*  
*Vishwakarma Institute of Technology*Pune, India  
aarya.nirgude21@vit.edu Vedant Mude   
*Department of Computer Engineering*  
*Vishwakarma Institute of Technology*Pune, India  
vedant.mude21@vit.edu

Tejashri Nirmal   
*Department of Computer Engineering*  
*Vishwakarma Institute of Technology*Pune, India  
tejashri.nirmal21@vit.edu Rutuj Nagrale   
*Department of Computer Engineering*  
*Vishwakarma Institute of Technology*Pune, India  
rutuj.nagrale21@vit.edu

*Abstract*— Images of historical documents often have characteristics, such as wrinkles, faint writing, stains, bleed-through ink, and other issues. These factors distort the text visibility and affect the performance of binarization. Preserving these document images aids future generations in learning about a variety of subjects. This article presents a new binarization approach for historic documents. This work uses bilateral and unsharp filtering, Otsu thresholding, histogram analysis, and k-means clustering for dark spot removal as part of a multi-step image enhancement process. Applying an intensity-based mask raises the quality of pixels above a set threshold. Furthermore, the method includes two additional refinements: a concluding sharpening phase and an enhancement of color contrast. The performance of proposed binarization approach is assessed using the Flesch Reading Ease Formula. The findings reveal that the algorithm achieved its highest readability score when applied to degraded documents, with an average readability score of 96.44. This suggests the algorithm's efficacy in enhancing the readability of noisy images, particularly in the context of degraded documents.

Keywords—Binarization, historical documents, Image filtering, Otsu thresholding, K-means clustering, Flesch Reading Ease Formula, image enhancement

# Introduction

Over thousands of years, a vast number of historical records with important information on literary arts and historical knowledge have been left behind. The historical document collection has seen significant deterioration, such as creases, faint text, stains, bleed-through ink, smeared ink, thin strokes, and degraded documents, due to years of storage. In recent years, a large number of researchers have become interested in the issue of document preservation. To safeguard printed paper documents against direct modification for consultation, exchange, and remote access, an effective method is to employ a document digitization system [1]. To increase accessibility and prevent physical copies from deteriorating, several libraries worldwide are digitizing historic documents. Document image processing algorithms are required to access the contents of these documents because document images cannot be read by machines in their raw form [2].

Noise reduction is the process of applying advanced image processing techniques to historic document photos to remove unwanted artifacts and interruptions. Certain methods are employed to reduce or remove the noises while preserving essential features; filters and denoising techniques are two examples of such algorithms. The goal is to enhance the text and image clarity while maintaining the original content integrity. This will make the historical document more readable and ensure its preservation for future usage [3].

In document images, it is common practice to map background pixels to white and foreground text pixels to black. Binarization is the process of transforming a multi-tone image into a bi-tonal image. Binarization can be used as a method of noise reduction to improve the readability of documents [4]. It is one of the preprocessing tasks that significantly affects other phases, such as feature extraction and recognition from document images, that call for a precise and high-quality foreground image [5]. Document image binarization is a crucial step in the pipeline for document image analysis and recognition that influences the outcomes of final recognition [6].

This paper studies rich literature on image enhancement and selects the most popular 5 filtering techniques. Through experimentation, this work obtains the optimized sequence of these techniques and proposes the system with the given sequence as sharpening filter, bilateral filter, k-means clustering, Otsu thresholding, and unsharp masking proves to be efficient for enhancement of historic document image. K-means clustering works well to eliminate dark areas, and histogram analysis makes it easier to understand how image intensity is distributed. After Otsu thresholding, an intensity-based mask is applied to further increase pixel values and establish an optimal threshold for binary conversion. The image details are sharpened once the color contrast has been adjusted. Together, these stages offer a comprehensive approach that enhances an advanced image-processing pipeline that can be used in a variety of disciplines. This method is applied to various categories of noisy images to obtain an average Reading Ease Score. A thorough study of the average scores of noisy images is presented further.

# Literature Review

The paper [1] presents an end-to-end trainable framework that carries out layout analysis, recognition, and character detection all at once. To improve recognition performance even further, it suggests a re-score technique that uses the implicit language model to forecast damaged characters. Annotations for layout, characters, and text lines have been added to the original dataset to aid in the study of Chinese historical writings.

The Binary Inpainting Network (BINet) is an autoencoder framework that uses binary inpainting to improve patch-based compression of static pictures. To facilitate parallel encoding and decoding, it restores interdependencies between nearby patches, doing away with the requirement for access to the original or rebuilt picture data. As mentioned in [7], at various compression levels, BINet significantly improves the compression quality.

An innovative strategy that has the advantage of simultaneously binarizing the image and reducing various types of noise is presented in the publication [8]. It converts an image that is noisy in grayscale to a much smaller low-noise binary image. The approach performs noticeably better when used on photos with similar text sizes and formats. As a result, official letters and document images from colleges or institutions greatly benefit from the algorithm.

A technique for removing random noise in several stages is presented in the publication [9]. Phase 1 involves scanning the complete text image top-down, then counting each connected black pixel region by going pixel by pixel through each column and row from left to right. These categories contain both noise-indicating sites and character/information locations. Phase 2 involves classifying the noise regions based on the number of black pixels in each numbered zone. To facilitate processing, phase 3 data is normalized to a range. Areas with pixels within the normalization range are used to represent data in phase 4. For more effective noise reduction, these stages are repeated a predetermined number of times.

The study [10] presents an algorithm that uses a step-by-step approach to image enhancement. Most document enhancement and cleaning procedures or binary picture conversions include a combination of intricate image processing techniques, increasing computing complexity and cost. The technology expedites the process by taking into account the distinct characteristics of the document images. Furthermore, the method is more user-friendly due to its iterated step structure.

There are four steps involved in using the algorithm: The process involves four steps: 1. determining the image's vector of parameters to be filtered; 2. applying a bilateral filter to filter the image; 3. dividing the image into its RGB components and binarizing each one using an approach based on Otsu's algorithm; and 4. selecting the RGB component that best preserves the document information in the foreground, which is the algorithm's final output [11].

The algorithm proposed in [3] does not provide a conclusion about histogram equalization; in fact, some of the images are deteriorated even more after histogram equalization. In terms of histogram equalization, the proposed approach provides conclusion for Triangle and Otsu thresholding. The work proposed in [6] works efficiently on a few images but is not tested on large number of images containing diverse noises while the proposed algorithm is tested on more than thousand images classified in 7 categories that are creases, faint text, stains, bleed-through ink, smeared ink, thin strokes and degraded image. This pipeline provides good result for images containing diverse noises while [5] fails to address this issue. The abundance of existing literature aids in the development of the suggested methodology, which improves historic document picture binarization outcomes.

# Methodology

For improved historic document image binarization, a combination of filtering approaches, including bilateral filter, k-means clustering, Otsu thresholding, sharpening filter, and unsharp masking, is suggested in this methodology. The workflow of the proposed approach is shown in Figure 1, and the following subsections provide a brief explanation of each stage.

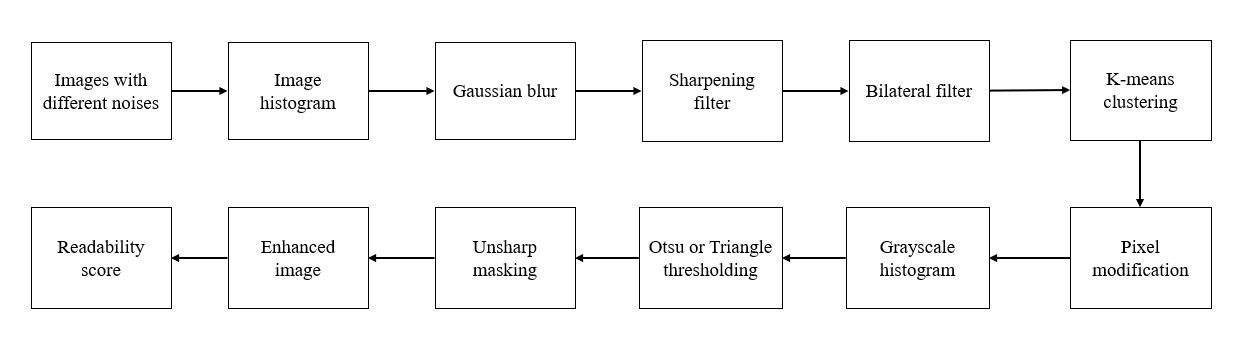


Figure 1. Proposed workflow

## INPUT DATA

The historical document photos used as input data come from a Kaggle repository [12] that has 1171 pictures with different kinds of noise in them. We have categorized those pictures into 7 categories creases, faint text, stains, bleed-through ink, smeared ink, thin strokes, and degraded. These 7 categories overlapped as a few noises altogether are present in the document images.

## IMAGE HISTOGRAM

A graphical representation of an image pixel intensity distribution is provided by an image histogram. It provides a graphical representation of the frequency of each intensity level, ranging from dark to bright. The horizontal axis represents the intensity values (i.e., pixel values in a grayscale image ranging from 0 to 255), and the vertical axis shows how frequently each intensity occurs. The knowledge and optimization of image properties during a range of image processing tasks can be aided by an understanding of the general contrast, brightness, and distribution of pixel values through the examination of an image histogram [13].

## GAUSSIAN BLUR

"Gaussian blur" is a convolutional image processing approach that includes using a Gaussian function to smooth an image [14]. This technique is often used to decrease noise and enhance photos. During the convolution process, the weighted average of the image pixel values is computed. The weights are determined by the values of a Gaussian kernel. A Gaussian kernel is a two-dimensional matrix where the core values gradually decline in a symmetric pattern, with the maximum weight allocated to them. The amount of blurring is determined by the standard deviation of the Gaussian function; a larger standard deviation results in a broader blur. Because Gaussian blur is computationally cheap and preserves edges to some extent, it is a versatile technique in image processing for applications such as pre-processing, background removal, and artistic effects.

## SHARPENING FILTER

Denoising and image sharpening are crucial in image processing. Image sharpening seeks to enhance edge slopes without producing halo artifacts, whereas an image-denoising algorithm seeks to minimize noise while preserving image edges. Image sharpening methods such as the Unsharp Mask filter improve edge slopes without causing halo artifacts. The Unsharp Mask does this by purposefully boosting high-frequency elements and emphasizing edges via a deftly designed blend of blurring and subtraction. This method provides a versatile tool for progressively improving overall image quality by minimizing unwanted visual artifacts and optimizing image details [15].

Let be the original image, and be a blurred version of the image, obtained by applying Gaussian blur with kernel size .

The Sharpened image is obtained by the process described below:

(1)

where, is the sharpened image, is the scaling factor that controls the strength of the sharpening effect, and the sub that represents the smoothened image from the original image represents the high-frequency components i.e., edges of the image.

## BILATERAL FILTER

A non-linear technique known as the bilateral filter can blur an image while preserving its sharp edges. Its ability to split an image into many scales without creating halo effects after manipulation makes it a popular choice for computational photography tasks such as tone mapping, denoising, style transfer, and relighting [16].

The bilateral filter is applied on the image mathematically as shown below:

(2)

where is the bilateral filter value at the spatial point of the image. The color or intensity of the pixel at position q in the image is represented by . The spatial neighborhood surrounding the pixel that the filter is applied to is denoted by Ω. The spatial weight is represented by , which is frequently defined as a Gaussian function depending on the spatial distance between and q. The range weight is represented by , which is frequently described as a Gaussian function based on the color or intensity difference between and .

## k-MEANS CLUSTERING

The image is divided into clusters using k-means clustering. Based on the given data set, k-means find new and undiscovered classes by segmenting the instance space into regions with related items. It is an iterative, unsupervised heuristic clustering method based on partitioning [17]. The dark spot clusters are discovered and treated individually to perhaps eliminate or minimize their influence on the image. Additionally, it aids in the improvement of text areas [18].

The process commences with the random initialization of cluster centroids, represented within the feature space of the image pixels. To assign each pixel to the nearest centroid, the algorithm computes the Euclidean distance between each pixel and every centroid This distance is expressed as:

(3)

Utilizing this distance metric, each pixel is assigned to the cluster associated with the closest centroid. Subsequently, the centroids are updated by calculating the mean of all pixels assigned to each cluster. Mathematically, the centroid update equation is defined as:

(4)

where, denotes the set of pixels assigned to the cluster . This iterative process continues until convergence, characterized by stabilized centroids or the fulfillment of predefined convergence criteria. The resulting clusters depict distinct regions in the image, facilitating tasks such as dark spot segmentation or text area enhancement. Through its iterative partitioning of pixel space and centroid re-computation, k-means clustering emerges as a versatile tool for uncovering underlying patterns within image data, thereby empowering a wide array of image processing applications.

## PIXEL MODIFICATION

Pixel modification refers to modifying the intensity of pixels in an image with a specific target pixel value by adding a specified increment to their intensity values while ensuring that the resulting pixel values remain within a valid range. This prevents overflow and underflow issues and ensures that the modified image remains in a valid format.

## OTSU THRESHOLDING

Among the popular global thresholding techniques, Otsu selects a threshold value by maximizing the metric known as the between-class variance, which is crucial to statistical discriminant analysis [19]. The Otsu thresholding method is commonly used for binarization, particularly in cases where the optimal threshold value needs to be determined automatically. Binarization involves converting a grayscale image into a binary image, where each pixel is classified as either foreground (usually represented as white) or background (usually represented as black).

More explanation is provided for the Otsu thresholding algorithm. Let's say an image has K different grayscale representations of its pixels (1, 2,…, K). Let N be the total number of pixels , and let be the number of pixels at level i. Likelihood of occurring for level i is determined by .

Let threshold T be used to separate the image into two classes, and . Pixels with levels [1,..., T] make , while pixels with levels [T + 1,..., K] make up . The cumulative probabilities represented by *(T)* and *(T)*, the mean levels by *(T* ) and *(T)*, and the variances of the classes by *T* and *T*, respectively. Equations (5) through (10) provide all of these values, as detailed in [20].

*(T) =*  (5)

*(T)= = 1­(T)* (6)

*(T)= =*  (7)

*(T)= =*  (8)

*T =*  (9)

*T =*  (10)

Let μ stand for the image mean level, *T* the between-class variance, and *T* the within-class variance. μ, *T*, and *T* are defined, respectively, by equations (11), (12) and (13).

*(T) (T) + (T) (T)*  (11)

*T = (T) + (T)*

(12)

*T = (T)T + (T)*  (13)

Equation (14) defines the cutoff that is reached by optimizing the between-class variance.

*= max{(T*)} (14)

This sum is equal to the cutoff found in equation (15) by lowering the within-class variances standard.

*= min{(T)}* (15)

## TRIANGLE THRESHOLDING

Triangle thresholding is one of the common thresholding techniques used to calculate the optimal thresholding value based on the histogram of the image. Let's denote the histogram as *H*(*x*), where *x* represents the intensity values. The peak of the histogram is denoted as *P* = (*xP*​,*yP*​), where *xP*​ is the intensity value and *yP*​ is the frequency at that intensity value [21].

Peak Detection:

(16)

Threshold Determination: The threshold value *T* is determined where the maximum distance occurs:

​ (17)

here, is the coordinate of the point on the histogram, is the perpendicular distance from a point to the line segment connecting the peak P.

## UNSHARP MASKING

Unsharp masking is a technique in image processing that improves image edges and features [22]. Sharpening, or unsharp masking, is the reverse of blurring, despite the name suggesting otherwise. Usually, a Gaussian blur is applied to do this, with the amount of blurring determined by the standard deviation. By averaging each pixel in the image with its nearby pixels, the blurring is produced.

(18)

where the blurred, original, and Gaussian blur pictures are represented, respectively, by , , and , with pixel coordinates .

Equation (19) shows that when the blurred image (mask) is subtracted from the original image, high-frequency elements like edges and features are highlighted in the resulting image.

(19)

where, is high pass filtered image.

Next, the high-pass filtered image is added back to the original image. This procedure increases the contrast of edges and little details, making them stand out more.

(20)

here, is a sharpened image.

Lastly, color contrast modifications contribute to improving the image brightness and visual attractiveness. A few adjustments that involve altering the saturation, color balance, or other features yield a more stunning and harmonious color representation. The final step of sharpening caps the method by adding crispness and refining the finer characteristics of the image. Each of these meticulously planned stages contributes to the overall goal of picture improvement and cleaning, resulting in a polished and eye-catching final output image.

## READABILITY EASE SCORE

The Flesch Reading Ease score [23] is a readability metric designed to quantify the ease with which a reader can comprehend a given text. It is calculated based on the average number of syllables per word and the average number of words per sentence in the text. The formula for the Flesch Reading Ease score is as follows:

(21)

here, is the Flesch readability which is ease score, is average words per sentence, is average syllables per word.

# Results and Discussions

The suggested technique produces an improved image by combining several filters, thresholding, clustering, and unsharp masking. The input images of various noise types are chosen from the dataset [12] depicted in the figures shown below. The input images are present with creases, faint text, stains, bleed-through ink, smeared ink, thin strokes, or are degraded due to the aging of paper. Following are the results of document images with various types of noises which undergo the process as described in the methodology to output visually appealing images.

Figures 2, 4, 6, 8, 10, 12 and 14 represents the input document images of categories: creases, faint text, stains, bleed-through ink, smeared ink, thin strokes and degraded image respectively and the corresponding cleaned images for each of these categories are given by figures 3, 5, 7, 9, 11, 13 and 15 respectively.

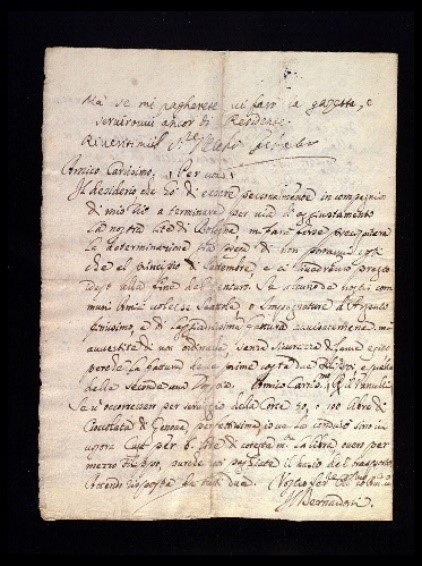


Figure 2. Creases

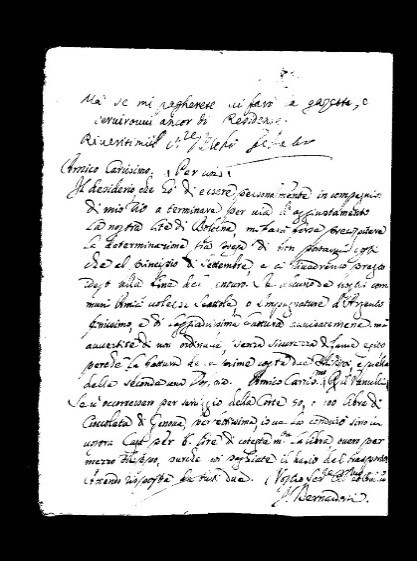


Figure 3. Cleaned crease noise

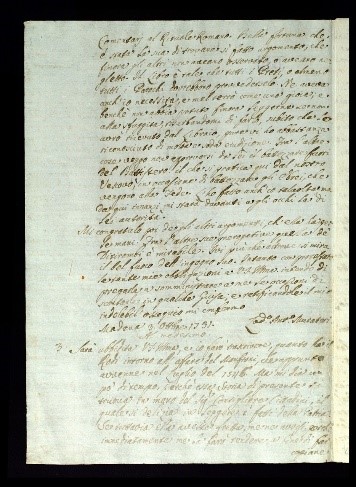


Figure 4. Faint text

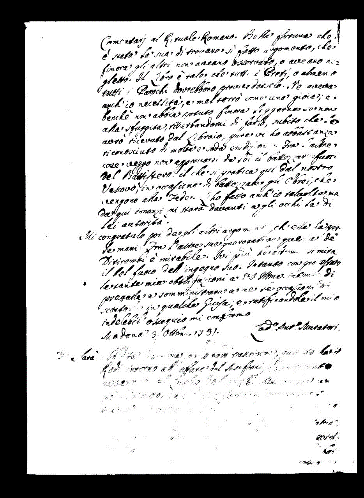


Figure 5. Cleaned faint text noise

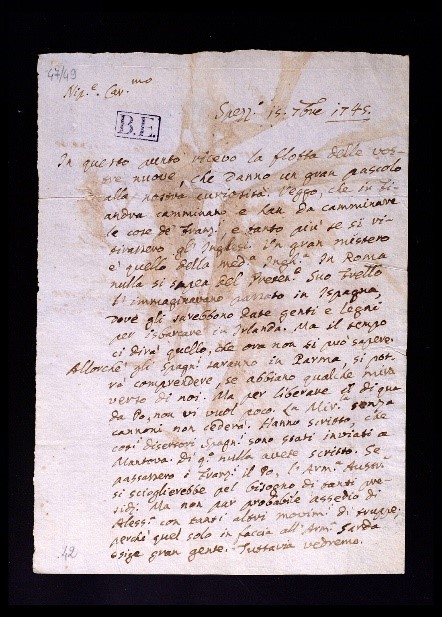


Figure 6. Stained image

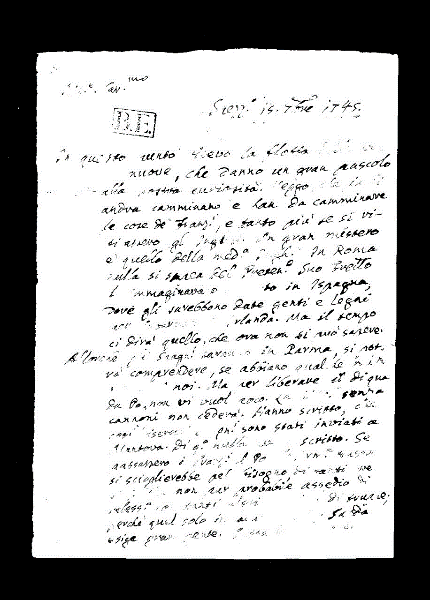


Figure 7. Stains cleaned from the image

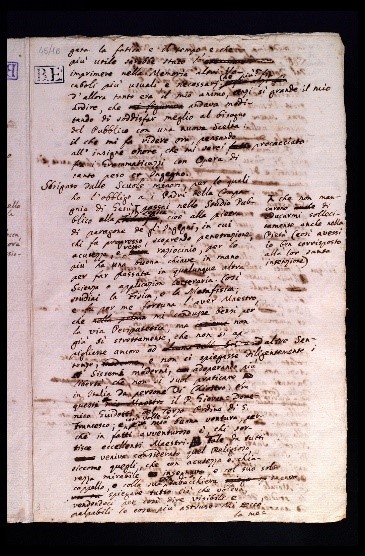


Figure 8. Bleed through ink image

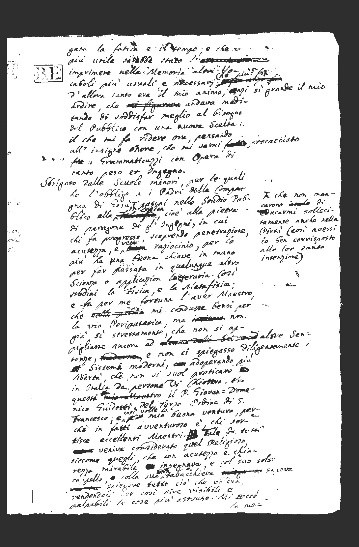


Figure 9. Cleaned bleed-through noise

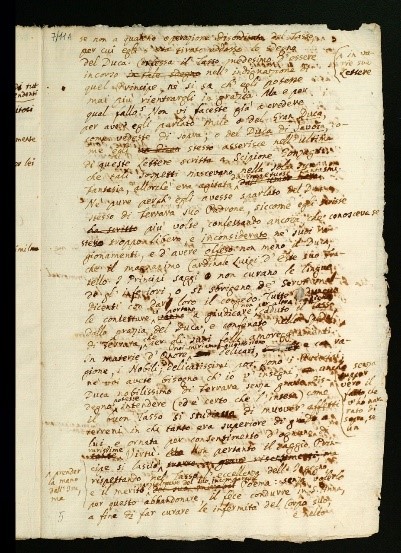


Figure 10. Smeared ink

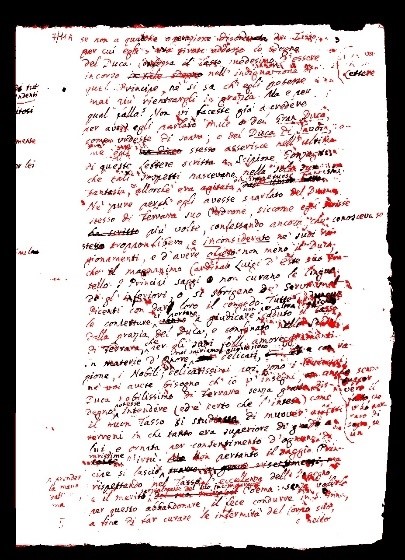


Figure 11. Enhanced image after removing smeared ink

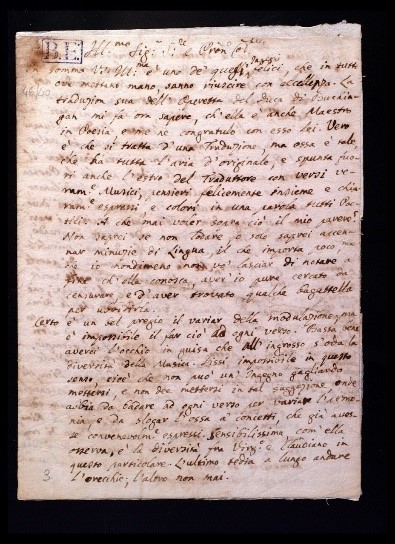


Figure 12. Thin strokes image

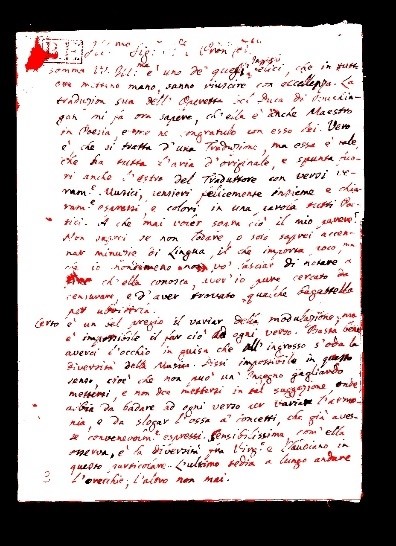


Figure 13. Cleaned thin strokes image

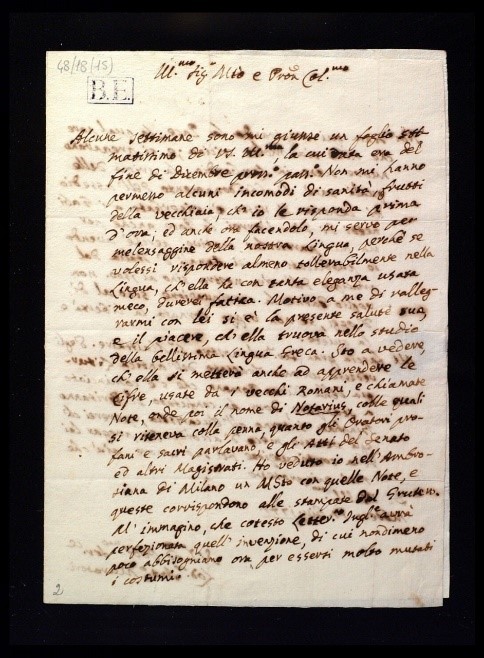


Figure 14. Degraded document image

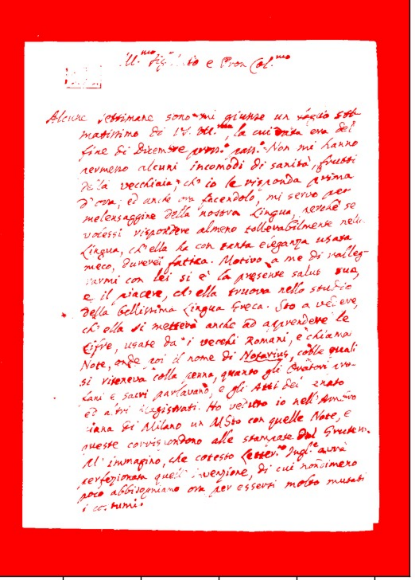


Figure 15. Enhanced document image

By applying the pipeline of the proposed binarization technique the resulting clean images are obtained for the different input historic documents of 7 categories based on the noises like creases, faint text, stains, bleed-through ink, smeared ink, thin strokes, and degraded image. The performance of the proposed binarization technique is evaluated using the Flesch Reading Ease score. The Average readability score for all output images is calculated. Table 1 summarizes the resultant average readability scores for various types of noisy images on application of the proposed binarization technique for historic document images.

## Table 1. The Resultant Average Readability Score for Different Noisy Images

|  |  |  |  |
| --- | --- | --- | --- |
| S. No. | Type of noise | No. of Images | Average Readability Score for the output image |
| 1 | Creases | 150 |  |
| 2 | Faint text | 100 |  |
| 3 | Stain | 320 |  |
| 4 | Bleed through ink | 200 |  |
| 5 | Smeared ink | 200 |  |
| 6 | Thin stroke | 150 |  |
| 7 | Degraded document | 400 |  |
| Average readability score | | | **87.77** |

# Conclusion

The current work uses a combination of various noise reduction and binarization approaches to improve the binarization of historic document images. For improved augmentation of the historical document images, a dependable and efficient technique has been found through the sequential application of unsharp filtering, bilateral filtering, k-means clustering, histogram analysis, Otsu thresholding, pixel modification, and final image creation. In addition to improving the noise reduction process, this comprehensive approach offers insightful information for future initiatives aimed at conserving and sharing cultural heritage through digitally preserved historical records. According to the proposed algorithm, the output image readability score is found to be higher than the original. As mentioned, the algorithm works best for degraded document images having an average readability score , as compared to other noises. The overall readability of the noisy historic documents after enhancement is .

Future scope comprises on natural language processing to intelligently digitize historic document collection. Additionally, future research on deep learning techniques, complex machine learning, semantic analysis integration, and natural language processing will all be included in the effort to intelligently digitize historical document collections. Additionally, the future scope includes creating an intuitive user interface to facilitate experimentation with the restoration of historic document images. This algorithm is efficient with diverse noise types in images but it should be tested on more images.

##### References

1. W. Ma, H. Zhang, L. Jin, S. Wu, J. Wang, and Y. Wang, “Joint Layout Analysis, Character Detection and Recognition for Historical Document Digitization,” in Proceedings of International Conference on Frontiers in Handwriting Recognition, ICFHR, Institute of Electrical and Electronics Engineers Inc., Sep. 2020, pp. 31–36.
2. R. Alaasam, B. Kurar, and J. El-Sana, “Layout analysis on challenging historical Arabic manuscripts using siamese network,” in Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, IEEE Computer Society, Sep. 2019, pp. 738–742.
3. G. Ganchimeg, “History Document Image Background Noise and Removal Methods,” International Journal of Knowledge Content Development & Technology, vol. 5, no. 2, Dec. 2015, pp. 11–24.
4. C. Tensmeyer and T. Martinez, “Historical Document Image Binarization: A Review,” SN Computer Science, vol. 1, no. 3. Springer, May 01, 2020.
5. Y. Akbari, A. S. Britto, S. Al-Maadeed, and L. S. Oliveira, “Binarization of degraded document images using convolutional neural networks based on predicted two-channel images,” in Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, IEEE Computer Society, Sep. 2019, pp. 973–978.
6. K. Ntirogiannis, B. Gatos, and I. Pratikakis, “Performance evaluation methodology for historical document image binarization,” IEEE Transactions on Image Processing, vol. 22, no. 2, 2013, pp. 595–609.
7. A. Nortje, W. Brink, H. A. Engelbrecht, and H. Kamper, “BINet: A binary inpainting network for deep patch-based image compression,” Signal Processing: Image Communication, vol. 92, 2013, p. 116-119.
8. A. Faranhmand, A. Sarrafzadeh, and J. Shanbehzadeh, “Noise Removal and Binarization of Scanned Document Images Using Clustering of Features,” International MultiConference of Engineers and Computer Scientists (IMECS), 2017, pp. 410-424.
9. B. M. Singh, K. K. Verma, and Mridula, “Noise Removal Technique for Document Images,” Journal of Multimedia Information System, vol. 10, no. 1, Mar. 2023, pp. 1–14.
10. E. Kavallieratou and H. Antonopoulou, “Cleaning and Enhancing Historical Document Images,” International Conference on Advanced Concepts for Intelligent Vision Systems, 2005, pp. 681-688.
11. M. Almeida, R. D. Lins, R. Bernardino, D. Jesus, and B. Lima, “A new binarization algorithm for historical documents,” Journal of Imaging, vol. 4, no. 2, 2018.
12. Kaggle Dataset, [Online]. Available at: <https://www.kaggle.com/datasets/vpippi/lam-dataset/data>
13. P. Daniel, R. Raju, and G. Neelima, “Image Segmentation by using Histogram Thresholding,” IJCSET, vol. 2, no. 1, 2012.
14. J. Flusser, S. Farokhi, C. Höschl, T. Suk, B. Zitová and M. Pedone, "Recognition of Images Degraded by Gaussian Blur," in IEEE Transactions on Image Processing, vol. 25, no. 2, Feb. 2016, pp. 790-806.
15. C. C. Pham and J. W. Jeon, “Efficient image sharpening and denoising using adaptive guided image filtering,” IET Image Processing, vol. 9, no. 1, Jan. 2015, pp. 71–79.
16. S. Paris, P. Kornprobst, J. Tumblin, and F. Durand, “Bilateral filtering: Theory and applications,” Foundations and Trends in Computer Graphics and Vision, vol. 4, no. 1, 2009, pp. 1–73.
17. S. M. Burney and H. Tariq, “K-Means Cluster Analysis for Image Segmentation,” International Journal of Computer Applications, vol. 96, no. 4, 2014.
18. N. Dhanachandra, K. Manglem, and Y. J. Chanu, “Image Segmentation Using K -means Clustering Algorithm and Subtractive Clustering Algorithm,” in Procedia Computer Science, vol. 54, 2015, pp. 764–771.
19. S. M. Elbayoumi Harb, N. A. M. Isa, and S. A. Salamah, “Improved image magnification algorithm based on Otsu thresholding,” Computers and Electrical Engineering, vol. 46, 2015, pp. 338–355.
20. X. Xu, S. Xu, L. Jin, and E. Song, “Characteristic analysis of Otsu threshold and its applications,” Pattern Recognition Letters, vol. 32, no. 7, May 2011, pp. 956–961.
21. S. D. Thepade and A. P. Bhalerao, "Machine Learning based Land Use Identification of Aerial Images with Fusion of Thepade SBTC and Triangle Thresholding," 2023 2nd International Conference for Innovation in Technology (INOCON), Bangalore, India, 2023, pp. 1-7.
22. Z. Shi, Y. Chen, E. Gavves, P. Mettes, and C. G. M. Snoek, “Unsharp mask guided filtering,” IEEE Transactions on Image Processing, vol. 30, 2021, pp. 7472–7485.
23. textstat library documentation [Online]. Available at: <https://pypi.org/project/textstat/>
24. S. Samer & A. Arafat, “Blank Background Image Lossless Compression Technique,” International Journal of Image Processing (IJIP), 2014, pp. 9-16.
25. A. Sekertekin, “A Survey on Global Thresholding Methods for Mapping Open Water Body Using Sentinel-2 Satellite Imagery and Normalized Difference Water Index,” Archives of Computational Methods in Engineering, 2021, pp. 1335–1347.