# Ancient Textual Restoration Using Deep Neural Networks

*Ali* Abbas Ali Alkhazraji[1,2]*, *Baheeja* Khudair[1] and *Asia* Mahdi Naser Alzubaidi[1]

[1]University of Karbala, College of Computer Science & Information Technology, Computer Science Department, Iraq
[2]College of Health and Medical Technology, University of Alkafeel, Iraq

**Abstract.** Ancient text restoration represents a critical area in computer science because it reflects an imagination about human life in early eras. Deep leaning plays a crucial role in AI last few years, specifically Generative Adversarial Networks (GANs), to regenerate and acclimatize old manuscripts that have suffered from the time effects, degradation, or deterioration. This work used Codex Sinaiticus dataset that preprocessed by encoding the dataset after that number and special character have been removed, new line symbol has been removed, tokenization process has been used to separate each word as an instance. Class target has been generated by removing character making it as a target and replacing it with special character. Using produces Generative Adversarial Networks (GANs), which consist of generator and discriminator inside in one learning framework. The generator part responsible for generating the missing text while the discriminator evaluates the generated text. But using an iteratively procedure these networks together collaboratively to provide a very sensitive reconstruction operations with the same format of ancient manuscripts, inscriptions and documents. Three prediction models used as proposed techniques for retrieving missing ancient texts are LSTM, RNN, and GAN and the results was validation accuracy 86%,92% and 98% respectively.

## 1 Introduction

Ancient texts are written records or documents that date back to ancient civilizations and societies. They serve as valuable windows into the past, offering insights into the history, culture, literature, religion, politics, science, and daily life of ancient peoples. These texts were typically inscribed on various materials such as papyrus, parchment, clay tablets, stone, bamboo slips, or even metal sheets, depending on the civilization and its technological advancements [1].

The restoration of ancient texts is of immense importance for various reasons, as it offers numerous benefits for scholars, historians, linguists, and society as a whole such as Preserving Cultural Heritage, Understanding History, Advancing Scholarship, Revealing Lost Knowledge, Resolving Historical Debates, Fostering Cultural Exchange, Correcting Biases and Misconceptions, Enhancing Language Studies, Enriching Literature and Arts, and Strengthening Cultural Identity.

These texts, “inscriptions”, are often damaged over the centuries, and illegible parts of the text must be restored. The main problem is to how restore the missing part of the ancient text to achieve the previous significance.

## 2 Related Work

The diligent endeavors of scholars specializing in the restoration of ancient texts afford us the opportunity to access a vast reservoir of knowledge that would otherwise remain obscured, thus enhancing our comprehension of historical events and the intellectual and aesthetic accomplishments of those who came before us. The restoration of these texts has a dual purpose: illuminating historical events and enhancing the continuous progression of human understanding while fostering a deeper appreciation for the rich and varied fabric of human history [2].

Languages serve as the foundation upon which human civilization is built, containing not just our methods of communication but also the very essence of our culture, thought, and historical development. They are the living reservoirs of the accumulated knowledge and expertise that has been accumulated through countless generations [3], [4].

* Corresponding author: ali.abbas.a@s.uokerbala.edu.iq

A group of scientists have worked in this field and obtained different results by using methods related to artificial intelligence

N. Shobha Rani et al. proposed a method for the restoration of text that has been marred by the presence of stain marks, ink seepages, and the effects of aging. These factors impede the enhancement of the document. This work addresses the challenges associated with tri-level semi-adaptive thresholding. However, the primary goal is to eliminate the damage that hinders the legibility of the letters. The proposed approach encompasses both pre- and post-enhancement phases, as well as three stages dedicated to the eradication of degradation. Pseudo-coloring uses a local thresholding technique, whereas level-wise degradation removal utilizes global thresholding. The process effectively eliminates ink and oil stains from palm leaves and DIBCO documents, while preserving the legibility of intricate language [5].

Cai et al. researched the recognition of ancient Chinese characters, encountering challenges such as suboptimal image quality and insufficient availability of annotated training data. The authors proposed a methodology utilizing Generative Adversarial Networks (GANs) and transfer learning as a means to facilitate the resolution of these issues. The Convolutional Neural Network (CNN) serves as the foundational framework for the discriminator, whereas the. The generator consists of an encoder-decoder architecture utilizing convolutional neural networks (CNNs). To evaluate the efficacy of the proposed method, namely TH-GAN, experimental investigations were conducted on two distinct tasks. The initial objective is the implementation of style transfer mapping for a diverse range of printed traditional Chinese character samples with several fonts. The subsequent objective involves the application of transfer learning to historical Chinese character samples through the incorporation of samples created by TH-GAN. The experimental findings demonstrate that the TH-GAN model, as presented in reference [6].

Ziran et al. conducted training on an initial model to identify frequently occurring words and hyphens inside web pages. The utilization of convolutional architectures, initially introduced for object recognition in natural photos, namely Faster R-CNN, involves training these structures to accurately identify and locate words included throughout the pages of the Bible that were printed by Johannes Gutenberg. Based on the model's prediction, a subsequent model is trained to identify specific key terms. The landmark words are commonly used and easily identifiable terms that are employed to align the image with an inaccurate page transcription [7].

Watanabe and colleagues (year) introduced a novel approach for character segmentation, which involves the utilization of a Fully Convolutional Network (FCN) followed by a post-processing phase applied to the network's outputs. The proposed methodology exhibits three key attributes: firstly, it possesses the capability to promptly analyze input images containing multiple lines of text, thereby obviating the requirement for text line segmentation. Secondly, it is not reliant on character recognition. Lastly, it demonstrates resilience in the face of variations in character size, character spacing, and overlapping. The accuracy of character segmentation in the context of Japanese historical handwritten government records has been reported to reach 95% in empirical testing [8].

A. Prusty et al. examined and elucidated the differentiation between instance segmentation and semantic segmentation. The identification of different text lines in Indic historical documents [22] is accomplished by employing a deep model that utilizes a Mask R-CNN with a ResNet-50 backbone. In contrast to alternative methodologies, this particular methodology encompasses a broader range of classifications, including Character line segment, Page border, Hole, Boundary line, Character component, and Physical degradation [9].

J. Pastor-Pellicer. Major Body Area (MBA) estimated text line for segmentation. The region of the corpus-baseline is MBA. Considered are two categories. Each pixel is designated as either a background, text block, or decorative by a sliding window that is centered on it. Text is distinguished using pixel tagging. To identify MBA pixels, a second CNN displays a sliding window focused on text blocks. The method is tested using data sets from Parzival and St. Gall. Only the proposed deep learning method consistently outperforms the other methods that were examined. Methods other than deep learning include text line segmentation using clustering algorithms, related component identification, and historical page analysis via anisotropic scale-space smoothing. The extraction of text lines using the Dynamic Multilayer Perceptron (DMLP) classifier produces comparable results [10].

K. Chen et al. suggested using a CNN with just one convolution layer. Super-pixel labeling, in which each pixel is classified as a backdrop, main text, decoration, or comment, is used to segment text lines. A super-pixel-centered patch serves as the categorization's input. Experiments have made use of several well-known datasets, including the George Washington, St. Gallen, and Parzival. The data shown here show a high level of precision. The proposed model, which produces accurate results, is contrasted to CRF and Multilayer Perceptron (MLP) as pixel classifiers for better understanding. Depending on the tested dataset, the suggested technique increases Intersection over Union (IU) from 1% to 4%, up to more than 90%, and superpixel labeling means accuracy [11].

## 3 The Structure Of Proposed System

To solve the mentioned problem, a proposed system was designed. Figure (1) explains the proposed system.

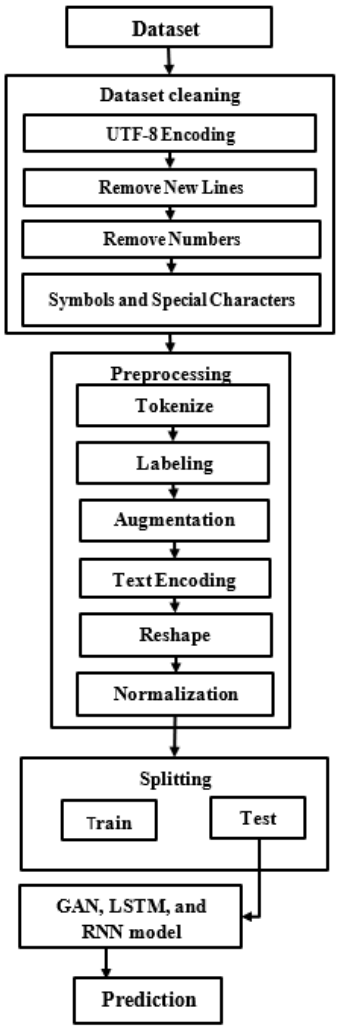


**Fig. 1.** Represents the proposed system

### 3.1 Dataset Collected

The Codex Sinaiticus, a manuscript dating back to the fourth century, is a significant artifact as it provides the earliest extant complete copy of the New Testament in the Christian Bible. The handwritten material is written in the Greek language. The New Testament is written in the native vernacular language of the time, known as koine, whereas the Old Testament is presented in the Septuagint form, which was embraced by early Greek-speaking Christians. Within the Codex, the textual content of both the Septuagint and the New Testament has undergone substantial annotation by a succession of early correctors. The Codex Sinaiticus holds significant relevance in the realm of reconstructing the original text of the Christian Bible, as well as in the fields of Bible history and Western book-making.

### 3.2 Dataset Cleaning

Dataset cleaning, also known as data cleaning or data cleansing, is an important stage in the data preparation process for any data-driven task, such as machine learning, data analysis, and database management. Here are some examples of frequent dataset-cleaning jobs and techniques:

#### *3.2.1 UTF-8 ENCODING*

The (UTF-8) is the standard Transformation Format 8-bit (UTF-8) used to employ representing text in multiple writing systems that are common in the world. UTF-8 encoding scheme exhibits unfixed length, such a character is represented by a list in the range of one to four bytes. it's important that commonly encountered characters, like ASCII, are typically encoded format a single byte. Also, UTF-8 has popped up as the general character encoding for the network of the World Wide Web and has Consolidated its position as the dominant standard method for text encoding in the computer science field.

#### *3.2.2 REMOVING THE NEW LINES, NUMBERS, AND SPECIAL CHARS*

In this step each empty line in the dataset is removed as a preprocess for the re-arranged dataset, to minimize the enhance the time of the dataset. regular expression substitution effective approach is to utilize a method used to remove numerical values and special characters in text raws of data. Regex or Regular expressions, that define a set of strings are general mathematical representations. In the same scenario, it can formulate the pattern of the regular expression that is represented by all characters that are not alphabetic or white spaces form, such as uppercase and lowercase letters in the range A to Z, same for spaces and tabs. as result, each occurrence of this specific form is changed or replaced by a null string, thereby reducing effectively removing each numeric values, special characters from the input text. The methods used in this preprocessing step introduce selective preserves only for the alphabetic symbols and spaces in the resulting text raw. This method is commonly used for text modification and analysis purposes.

### 3.3 Preprocessing

In this step, some of the stages had been applied to the dataset as a preprocessing to meet the model. These stages are:

#### *3.3.1 TOKENIZE*

To be utilized in the domain of the natural language processing (NLP), the Tokenization technique is one of the fundamental text preparation techniques that is commonly used. to divide an input text into smaller contents are known as tokens. the tokens can include phrases, words, and symbols, each of them representing significant aspects of input text and taking a main role in various natural language processing (NLP) operations like text analysis, sentiment, or machine translation, and text processes. through the first phases, are often provided, which consists of each one of input text in all its forms. Text tokenization is a process that handles place at multiple separated levels of text, in general, most prevalent can the sentence or word level. in the lexical level, the raw text is a general method split into disconnected units known by words, such as each word is represented by independent token. in the level of Word, tokenization is of very importance in natural language processing (NLP) processes like as part of speech fields, known recognition, and machine translation applications.

#### *3.3.2 LABELING*

The labeling process has been done by taking each token and choosing a random location within the length of that token. This location will be replaced with a missing character and the character with this location will be the target class for this token.

#### *3.3.3 AUGMENTATION*

The tokens in whole the dataset will be computed in order to know each token how many times repeated in the dataset the token with twenty times repetitions will be duplicated many times until the repetition number is equal to twenty.

#### *3.3.4 TEXT ENCODING*

In the section of deep learning, the idea of text encoding is represented by the method of converting randomly strcut of text data into a numerical form, which is facilitated using as input in all deep networks and machine learning models. The using of encoding is a very important process and also a necessity for deep learning models to handle numerical input layers.

encoding of Text data is a method that is commonly used in the parts of natural language processing (NLP) containing most of the techniques, such as the method of word embeddings as Word2Vec methods or BOW .

The word embeddings can the represented by words as an array, vectors in high-dimensional spaces or one-dimensional vectors, allowing for the consider as semantic relationships. Also, the one-hot, or categorical encoding is a commonly employed method used in the various of representing words as 0,1 values vector. those methods are encodings that improve of the potential of deep learning models to improve and generate forms, and semantic forms from text data, also enabling methods to analyze and discipline the human languages, and one-hot encoding method has been used in this research as a preprocessing method to generate the target of missing characters.

#### *3.3.5 RESHAPE*

The term “reshape” refers to the process of modifying the dimensions or structure of a data tensor, while maintaining the integrity of its underlying data. Implementing this process is of utmost importance to modify the data according to the specific requirements of the neural network layers. One common procedure in image data processing is to fold a 2D image matrix into a 1D vector. This transformation is often necessary as input to the fully linked layer. Process reconfiguration plays a crucial role in achieving consistency across multiple levels and facilitating the smooth flow of data within the network. This allows deep learning models to properly understand complex data patterns.

#### *3.3.6 NORMALIZATION*

In the era of deep learning networks and machine learning normalization is an important step among the steps of data preprocessing. The objective of the Normalization process is to adjust and scale the raws of the input features, so as to ensure accurate results features are insured within a specific range. sometimes the range is defined as containing random values by the range of 0 to 1, or as an alternative, as including a mean of raw feature 0 and a default of deviation by 1. The step of normalization is of in general critical in the model training phase, such as it enhances such that each of the inputs is raw and makes a good contribution. Its target of mitigate the range of the larger scale of all features in the all AI methods learning workflow and enhance the time of convergence during model training. The using of regularization methods is of general importance in the section of deep neural networks and also has a vital role in enhancing the work processes of the training and enhancing the ability of the model to produce perfect generalized to generated data. there are two commonly employed methods of normalization techniques such as the Z-score method, and the scaling method. There are multiple strategies are used to enhance the text data to ensure that it applies to ranges pre-determined.

### 3.4 GENERATIVE ADVERSARIAL NETWORK MODEL

Generative Adversarial Network (GAN) is a deep learning system consisting of two neural networks, the generator and the discriminative network, which participate in the adversarial learning process. The generator is responsible for producing synthetic data, such as images or text, to closely resemble the original data. Conversely, the primary function of the discriminator is to distinguish between real data and data generated by the generator. Using iterative training, the generator gradually gains the ability to generate data that poses a greater challenge for the discriminator to distinguish from the original data. Generative Adversarial Networks (GANs) have garnered significant traction across several areas owing to their capacity to provide innovative and superior-quality data, hence finding applications in fields such as art, image processing, and even drug development [12], figure (2) Illustrates GAN Architecture.

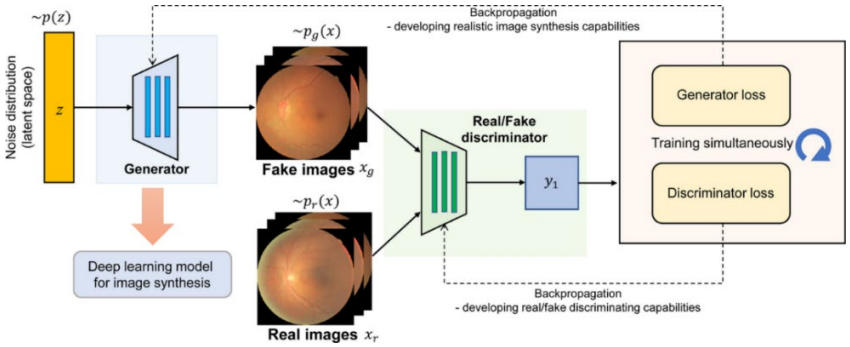


**Fig. 2.** An illustration of a basic architecture of GAN (vanilla GAN) for retinal image synthesis.

## 4 Evaluation Measures

When constructing a machine learning model, the evaluation of performance and efficiency is crucial. For the machine learning model to be trustworthy, an assessment method must be selected that is proportional with the model's work. Frequently, while assessing machine learning models, many scales are employed to guarantee accurate evaluation. In machine learning, there are three primary types of evaluation measures: those used to assess classification and clustering tasks. Classification tasks may be evaluated using a variety of metrics, including:

- Accuracy measure.
- Confusion Matrix (CM ).
- Recall.
- Precision.
- F1-score.

## 5 Result

The dataset has been trained and tested three prediction models LSTM, RNN, and GAN and the best results achieved in GAN model as below:

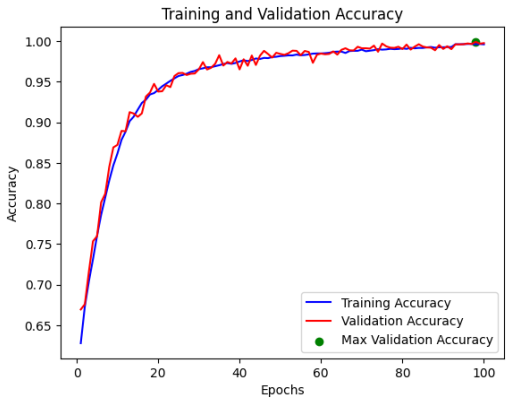


**Fig. 3.** The training and validation accuracy of GAN model.

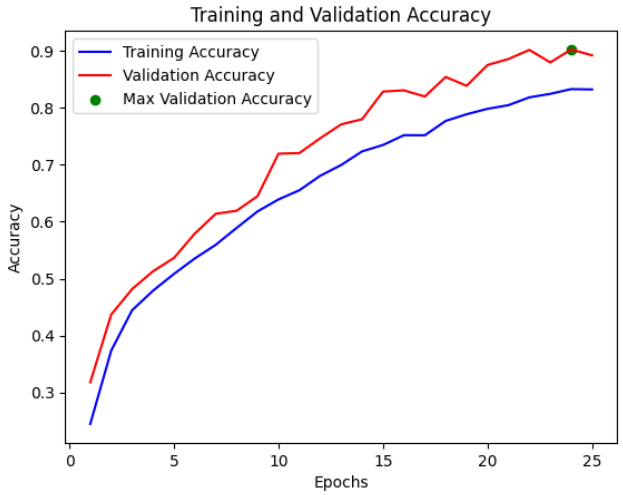


**Fig. 4.** The training and validation accuracy of LSTM model**.**

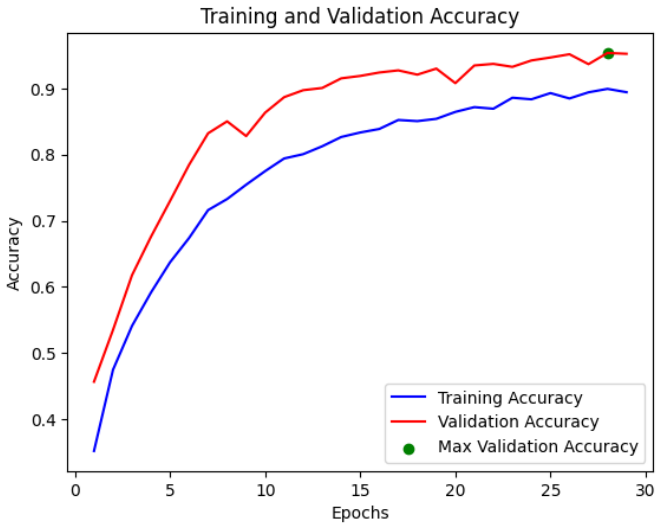


**Fig. 5.** The training and validation accuracy of RNN model.

## 5 Conclusion

Within a competitive learning context, this method employs Generative Adversarial Networks (GANs), which consist of a generator and a discriminator. The generator component is used to recreate missing or damaged text, whilst the discriminator component determines the authenticity of the generated material. These networks collaborate iteratively to build highly exact and contextually consistent reconstructions of ancient manuscripts, inscriptions, and documents. This development in textual restoration not only provides a limited perspective into historical events, but it also paves the door for the recovery of precious historical, literary, and cultural relics that were previously believed irreparable.

## References

1. V. Romero, A. H. Toselli, L. Rodríguez, and E. Vidal, “Computer assisted transcription for ancient text images,” Springer-Verlag Berlin Heidelb., vol. 4633 LNCS, pp. 1182–1193, 2007, doi: 10.1007/978-3-540-74260-9_105.
2. G. Carpenè, D. Negrini, B. M. Henry, M. Montagnana, and G. Lippi, “Homocysteine in coronavirus disease (COVID-19): a systematic literature review,” Diagnosis, vol. 9, no. 3, pp. 306–310, 2022, doi: 10.1515/dx-2022-0042.
3. M. Llamas, M. L. Garo, and L. Giovanella, “Low free-T3 serum levels and prognosis of COVID-19: Systematic review and meta-analysis,” Clin. Chem. Lab. Med., vol. 59, no. 12, pp. 1906–1913, 2021, doi: 10.1515/cclm-2021-0805.
4. A. Savelyev and M. Robbeets, “Bayesian phylolinguistics infers the internal structure and the time-depth of the Turkic language family,” J. Lang. Evol., vol. 5, no. 1, pp. 39–53, 2020, doi: 10.1093/jole/lzz010.
5. N. Shobha Rani, B. J. Bipin Nair, M. Chandrajith, G. Hemantha Kumar, and J. Fortuny, “Restoration of deteriorated text sections in ancient document images using a tri-level semi-adaptive thresholding technique,” Automatika, vol. 63, no. 2, pp. 378–398, 2022, doi: 10.1080/00051144.2022.2042462.
6. Cai, J.; Peng, L.; Tang, Y.; Liu, C.; Li, P. TH-GAN: Generative Adversarial Network Based Transfer Learning for Historical Chinese Character Recognition. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, 20–25 September 2019; pp. 178–183.
7. Ziran, Z.; Pic, X.; Innocenti, S.U.; Mugnai, D.; Marinai, S. Text alignment in early printed books combining deep learning and dynamic programming. Pattern Recognit. Lett. 2020, 133, 109–115.
8. Watanabe, K.; Takahashi, S.; Kamaya, Y.; Yamada, M.; Mekada, Y.; Hasegawa, J.; Miyazaki, S. Japanese Character Segmentation for Historical Handwritten Official Documents Using Fully Convolutional Networks. In Proceedings of the 2019 International Conference on Document Analysis and Recognition, ICDAR2019, Sydney, Australia, 20–25 September 2019; pp. 934–940.
9. Prusty, A.; Aitha, S.; Trivedi, A.; Sarvadevabhatla, S.R.K. Indiscapes: Instance Segmentation Networks for Layout Parsing of Historical Indic Manuscripts. In Proceedings of the 2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, 20–25 September 2019; pp. 999–1006.
10. Pastor-Pellicer, J.; Afzal, M.Z.; Liwicki, M.; Castro-Bleda, M.J. Complete system for text line extraction using convolutional neural networks and watershed transform. In Proceedings of the 2016 12th IAPR Workshop on Document Analysis Systems (DAS), Santorini, Greece, 11–14 April 2016; pp. 30–35.
11. Chen, K.; Seuret, M.; Hennebert, J.; Ingold, R. Convolutional neural networks for page segmentation of historical document images. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; Volume 1, pp. 965–970.
12. You, A., Kim, J. K., Ryu, I. H., & Yoo, T. K. (2022). Application of generative adversarial networks (GAN) for ophthalmology image domains: a survey. Eye and Vision, 9(1), 1-19.7] ] T. Murphy, "Evaluate open source risks," http://www. ftponline. com/wss/2002_10/online/tmurphy/, retrieved, vol. 19, p. 2012, 2001.