EAI Endorsed Transactions
on AI and Robotics

Research Article EAI.EU

A Deep Learning Based Optical Character Recognition Model for Old Turkic

Seyed Hossein Taheri [1,*], Houman Kosarirad [2], Isabel Adrover Gallego [3] and Nedasadat Taheri [4]

[1] Department of Convergent Sciences and Technologies, Islamic Azad University, Science and Research Branch, Tehran, Iran
[2] School of Computing, University of Nebraska–Lincoln, Lincoln, NE, USA
[3] School of Computing, University of Nebraska–Lincoln, Lincoln, NE, USA
[4] School of Computing, University of Nebraska–Lincoln, Lincoln, NE, USA

Abstract

This study presents the development and evaluation of a deep learning-based optical character recognition (OCR) model specifically designed for recognizing Old Turkic script. Utilizing a convolutional neural network (CNN), the project aimed to achieve high classification accuracy across a dataset comprising 38 distinct Old Turkic characters. To enhance the model's robustness and generalization capabilities, sophisticated data augmentation techniques were employed, generating 760 augmented images from the original 38 characters. The model was rigorously trained and validated, achieving an overall accuracy of 96.34%. Evaluation metrics such as precision, recall, and F1-scores were systematically analyzed, showing superior performance in most classes while identifying areas for further optimization. The results underscore the effectiveness of CNN architectures in specialized OCR tasks, demonstrating their potential in preserving and digitizing historical scripts. This study not only advances the field of document analysis and OCR but also contributes to the digital preservation and accessibility of ancient scripts.

Keywords: Deep learning, optical character recognition, convolutional neural network

Received on 17 January 2025, accepted on 01 April 2025, published on 10 April 2025

doi: 10.4108/airo.8460

1. Introduction

The Old Turkic script, an integral component of Central Asian Turkic cultural heritage, has long presented unique challenges in the digitization and preservation of historical texts. Dating from the 8th to the 13th centuries, this script is primarily associated with monumental inscriptions and holds significant importance for understanding Turkic history, culture, and linguistic evolution. The angular, runic-like characters, often found in stone inscriptions and manuscripts, require sophisticated recognition techniques for digital conversion, as these ancient scripts differ markedly from contemporary texts in orthography and preservation condition [1].

Advancements in OCR technology have generally focused on modern scripts, often neglecting the peculiarities of ancient alphabets like Old Turkic. However, the adoption of deep learning techniques, particularly CNNs, has started to transform this field. These modern approaches not only offer enhanced accuracy in character recognition but also extend the accessibility of these invaluable texts to a broader audience, including researchers, historians, and the general public.

This project delves into the development of a specialized deep learning-based OCR model tailored to the Old Turkic script. By leveraging a dataset comprising 38 distinct characters, the model benefits from rigorous training regimes and sophisticated data augmentation techniques

[*] Corresponding author. Email: hkosarirad2@huskers.unl.edu

aimed at maximizing recognition accuracy and model robustness. This introduction to the field of digital humanities highlights several pioneering efforts that have successfully utilized deep learning to bridge the gap between historical and modern languages. Through this project, we aim to contribute significantly to the preservation of Turkic cultural heritage and set a benchmark for future endeavours in the digitization of ancient scripts. The ensuing sections will detail the methodologies employed, including data collection, preprocessing, model architecture, data augmentation, and evaluation, culminating in a discussion of the results and their implications for our future research and work.

2. Background

The Old Turkic script, also known as Orkhon script, is a critical part of the cultural and historical heritage of various Turkic peoples of Central Asia, used from the 8th to the 13th centuries. It is named after the Orkhon Valley in Mongolia, where the earliest known inscriptions in this script were discovered in 1889. Predominantly employed for monumental inscriptions, this script is essential for understanding the Turkic history, culture, and linguistic evolution. The angular, runic-like characters of the Old Turkic script, found in a range of historical documents from stone inscriptions to manuscripts, present unique challenges for recognition and digitization due to its distinctive features and variations. The digitization of historical texts is a vital process in the preservation and accessibility of cultural heritage. Traditional OCR systems often struggle with ancient scripts, which vary significantly from modern texts in terms of orthography and medium preservation. Developing a deep learning-based OCR model for the Old Turkic script not only tackles these technical challenges but also ensures that these invaluable texts are preserved and made accessible to a broader audience, including researchers, historians, and the general public. In the broader field of digital humanities, several pioneering projects have demonstrated the potential of deep learning in handling ancient scripts. For instance, projects like [2] and [3] have successfully bridged the gap between historical and modern languages, enhancing accessibility and understanding. Similarly, [4] and [5] have shown significant advancements in recognizing and digitizing scripts with complex characteristics.

Comparative studies have also evaluated traditional OCR methods such as Tesseract and ABBYY FineReader against deep learning-based approaches, finding that CNN models significantly outperform traditional OCR systems, especially when handling degraded and ancient texts [6]. For instance, a comparative analysis by Patel et al. demonstrated that deep learning-based methods achieved an accuracy of up to 93% on historical texts, whereas traditional OCR tools like Tesseract managed accuracy levels of around 40% [6].

Recent advancements in Transformer-based OCR models such as TrOCR have also been highlighted in the literature, showcasing superior performance due to their self-attention mechanisms, enabling them to capture long-range dependencies in textual data more effectively than CNN-based models [7]. Recognizing historical scripts also introduces additional challenges, such as character degradation, irregular spacing, and limited datasets, which have been extensively discussed in the literature. Studies underline the potential of hybrid models combining CNNs and Transformers to further enhance OCR performance in these challenging scenarios [8].

Recent advancements in deep learning have significantly reshaped the OCR field, particularly with the emergence of Transformer-based models such as TrOCR and Vision Transformers (ViTs). Transformer-based OCR models rely on self-attention mechanisms rather than convolutional layers, enabling them to effectively capture contextual relationships and long-range dependencies within textual data, which significantly enhances recognition accuracy, especially in challenging scenarios like historical or degraded scripts.

TrOCR, introduced by Li et al. (2021), combines a Vision Transformer-based encoder with a Transformer decoder to achieve state-of-the-art OCR accuracy on multiple benchmark datasets. Li et al. demonstrated that TrOCR achieved substantially improved accuracy on both printed and handwritten texts compared to conventional CNN-based OCR models, particularly in scenarios involving complex text layouts or degradation [9].

Similarly, Vision Transformers (ViTs), originally proposed by Dosovitskiy et al. (2021), have also shown superior performance in OCR tasks, leveraging self-attention mechanisms to process entire images as sequences of patches. ViTs have proven highly effective in scenarios that require modeling global context, a notable limitation of traditional CNN architectures which rely predominantly on local feature extraction [10].

Nevertheless, CNN-based architectures, as used in our study, remain widely adopted due to their relatively lower computational cost and ease of implementation, making them particularly suitable for scenarios with constrained computational resources or limited dataset sizes. The CNN-based approach employed in this study leverages these strengths, offering an efficient solution while acknowledging that Transformer-based approaches may represent a promising direction for future research in OCR of historical scripts.

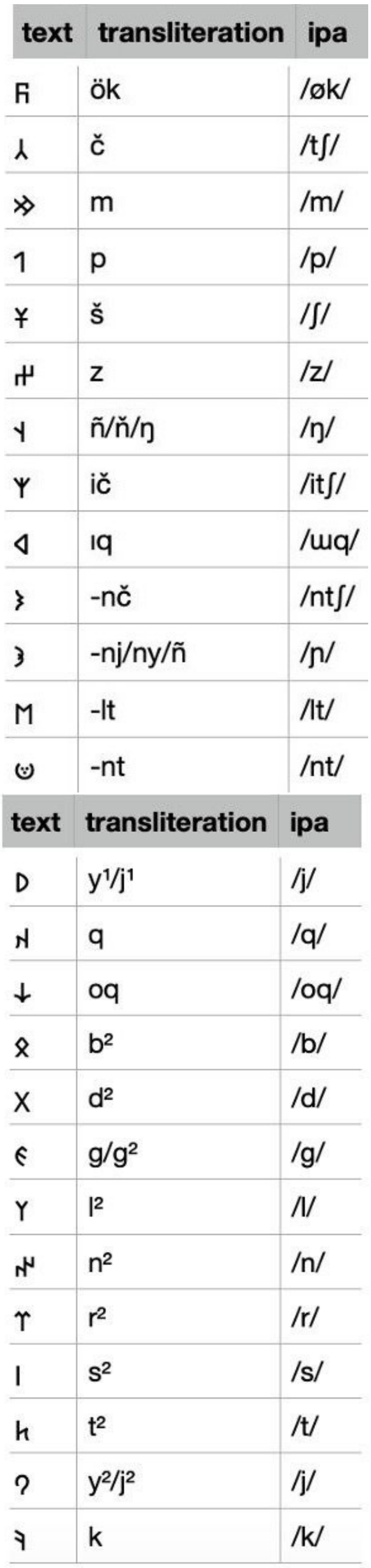
On the other hand, Historical script recognition presents unique challenges compared to contemporary texts. Ancient documents often suffer from physical degradation, making characters difficult to distinguish clearly. Furthermore, historical manuscripts frequently exhibit irregular spacing and inconsistent alignment, complicating the segmentation process in OCR systems. Additionally, the limited availability of annotated datasets for ancient scripts significantly hampers the training of robust OCR models. According to Fischer et al. (2020), character degradation and irregularities are among the most significant hurdles in historical text digitization, requiring specialized preprocessing methods and advanced recognition algorithms to mitigate errors [11]. Another study by Clanuwat et al. (2019) emphasizes these challenges, demonstrating how deep learning models can partially address dataset scarcity through transfer learning and synthetic data augmentation strategies [12]. Given these challenges, deep learning-based models, particularly CNN and Transformer architectures, are widely recommended due to their inherent ability to learn complex patterns from limited and noisy datasets.

The collective success of these technologies in various linguistic contexts illustrates the adaptability and potential of deep learning models to revolutionize the field of historical text digitization. By adopting and refining these technologies for the Old Turkic script, this project aims not only to contribute to the preservation of Turkic cultural heritage but also to set a precedent for future research and development in the digitization of other ancient scripts. This work will facilitate new research opportunities in Turkic studies and could potentially lead to breakthroughs in the study of other culturally significant scripts.

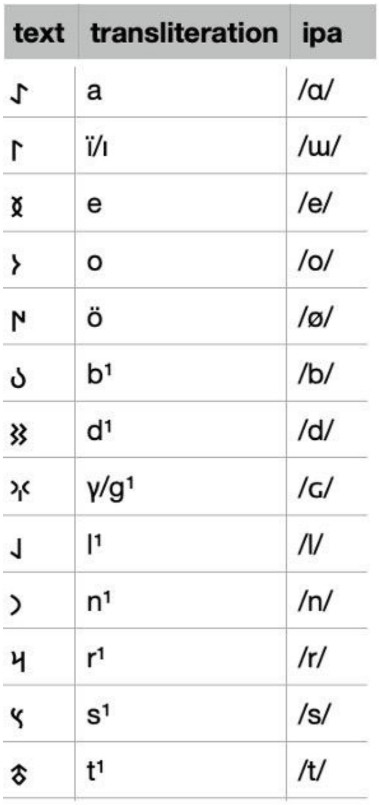
3. Methodology

3.1. Data Collection

The data collection for the study involved curating a dataset of Old Turkic script characters from Wikipedia [13], which provided a reliable and accessible source for ancient script images and corresponding textual information as shown in Fig.1 . The gathered data encompasses images of specific characters, each labeled with unique text, transliteration, and International Phonetic Alphabet (IPA) annotations. This col- lection was systematically organized into a CSV file format, featuring columns for ‘image id‘, ‘text‘, ‘transliteration‘, and ‘ipa‘. For instance, the entry ‘image1.png‘ corresponds to the first label with a transliteration of ”a” and an IPA notation of /a/. This methodical compilation of 38 character images, along with their linguistic attributes, resulted in a structured dataset that facilitates automated processing and analysis, serving as the foundational corpus for training the neural network model. This process not only preserved the linguistic integrity of the characters but also ensured a standardized format for efficient machine-learning applications.



text	transliteration	ipa
𐰜	ök	/øk/
𐰲	č	/tʃ/
𐰢	m	/m/
𐰯	p	/p/
𐱁	š	/ʃ/
𐰔	z	/z/
𐰭	ñ/ň/ŋ	/ŋ/
𐰱	ič	/itʃ/
𐰶	ıq	/ɯq/
𐰨	-nč	/ntʃ/
𐰪	-nj/ny/ñ	/ɲ/
𐰡	-lt	/lt/
𐰦	-nt	/nt/
𐰖	y^1/j^1	/j/
𐰴	q	/q/
𐰸	oq	/oq/
𐰋	b^2	/b/
𐰓	d^2	/d/
𐰏	g/g^2	/g/
𐰠	l^2	/l/
𐰤	n^2	/n/
𐰼	r^2	/r/
𐰾	s^2	/s/
𐱅	t^2	/t/
𐰘	y^2/j^2	/j/
𐰚	k	/k/



text	transliteration	ipa
𐰀	a	/ɑ/
𐰃	ï/ı	/ɯ/
𐰅	e	/e/
𐰆	o	/o/
𐰇	ö	/ø/
𐰉	b^1	/b/
𐰑	d^1	/d/
𐰍	γ/g^1	/ɢ/
𐰞	l^1	/l/
𐰣	n^1	/n/
𐰺	r^1	/r/
𐰽	s^1	/s/
𐱃	t^1	/t/

Figure 1. Characters

3.2. Data Preprocessing

In the data preprocessing phase of the study, images are uniformly resized to 32x32 pixels and normalized after extracting the green channel to prepare them for a CNN. The dataset, consisting of image filenames and corresponding labels, is loaded and categorized from a CSV file. The choice of 32x32 pixels was initially selected for computational efficiency and faster training times, in alignment with previous studies demonstrating sufficient accuracy for similar tasks with simple character shapes [16]. Non-loadable images are excluded to maintain data integrity. To enhance model robustness and prevent overfitting, the dataset undergoes augmentation through the Image Data Generator from Keras, which applies random rotations, shifts, and zooms, effectively increasing the dataset size by generating 20 variations per image. Subsequently, the dataset is split into training and validation sets, ensuring the model's performance is reliably evaluated on unseen data during the training phases. The decision to utilize only the green channel from the RGB images was based on empirical observations from previous studies indicating that the green channel typically provides superior contrast and image clarity, especially when processing digitized documents and manuscripts captured under varying lighting conditions [14]. Compared to grayscale conversion, the green channel often retains higher contrast details due to its proximity to human visual sensitivity peaks, resulting in clearer feature distinctions for the CNN model [15]. Additionally, prior research by Yang et al. (2019) demonstrated improved recognition performance when using the green channel specifically, especially for historical document images, due to better noise suppression and feature preservation compared to full RGB inputs or grayscale conversions [16].

By incorporating this preprocessing step, we aimed to maximize the distinguishability of character features, which is crucial for improving the CNN's accuracy in classifying Old Turkic script.

3.3. CNN Model

The construction of the CNN model involves a sequential architecture configured through TensorFlow's Keras API to effectively learn from the prepared image data (Fig.2). The model comprises multiple layers designed to capture varying levels of abstraction in the image data: three convolutional layers each followed by max-pooling layers to extract and downsample features, and a flattening step that transitions the 2D feature maps into a 1D feature vector. Each convolutional layer utilizes 32, 64, and 128 filters, respectively, with a kernel size of 3x3 and 'relu' activation, optimizing the extraction of detailed features. Post-flattening, two dense layers with 'relu' activation precede the output layer, which employs a softmax activation function to output the probabilities for each class. The model is compiled using the Adam optimizer and categorical crossentropy loss function to fine-tune the parameters and minimize classification errors. This structured layering and choice of hyperparameters aim to balance the model's capacity and computational efficiency, facilitating effective learning and prediction on the image classification task.
The CNN architecture adopted in this study was deliberately designed to balance model complexity with computational efficiency. While deeper CNN architectures such as ResNet or advanced Transformer-based models like Vision Transformers (ViTs) offer superior representation power, they require significantly greater computational resources, larger datasets, and longer training times [20,21]. Given the limited dataset size available for the Old Turkic script and resource constraints, a relatively simpler CNN with three convolutional layers was chosen. This simpler architecture effectively reduces the risk of overfitting on our limited dataset, ensures faster training and inference times, and remains computationally manageable for practical deployment scenarios [17]. Moreover, previous studies have demonstrated that shallow CNN architectures can still achieve very high accuracy in similar OCR tasks involving historical scripts, provided that careful data augmentation and preprocessing are employed [17]. Therefore, our selected model represents an optimized compromise between complexity, computational feasibility, and practical OCR performance. Future research may explore deeper or hybrid

architectures, such as CNN-Transformer combinations, to potentially achieve even higher recognition accuracy.

3.4. Data Augmentation

The data augmentation process employed in the study significantly enriches the dataset by generating varied instances of each original image to enhance model robustness and generalization. Utilizing the Image Data Generator from Keras, each of the 38 original character images undergoes 20 distinct transformations, resulting in a total of 760 augmented images. Sample of this augmentation has been shown in Fig.3 These transformations include random rotations up to 10 degrees, width and height shifts by 0.1, shear adjustments of 0.1, and zoom variations of 0.1, all of which are designed to mimic real-world discrepancies in image capture conditions. The Fill model parameter is set to 'nearest' to maintain the integrity of transformed pixels. This augmentation strategy not only diversifies the training data but also effectively increases the dataset size, providing the CNN with a broader range of features to learn from, thereby enhancing the model's ability to perform accurately on unseen data.

3.5. Model Evaluation

In evaluating the performance of the CNN model, the study employs a set of metrics to assess accuracy and loss during both the training and validation phases. Accuracy, the primary metric, measures the proportion of correctly predicted labels against the total number of samples, providing a direct indication of the model's classification performance. Loss, measured by categorical cross-entropy, quantifies the disparity between the predicted probabilities and the actual labels, offering insight into the model's predictive confidence and error magnitude. These metrics are monitored throughout the training process to observe trends and potential overfitting. Validation metrics are particularly crucial as they reflect the model's capability to generalize to new, unseen data, beyond the training set. By analyzing these metrics, the study not only confirms the effectiveness of the model's learning but also identifies opportunities for further optimization to enhance predictive accuracy and reduce loss.

To evaluate the effectiveness of the proposed CNN model, its performance was directly compared against well-established OCR techniques, including traditional OCR systems such as Tesseract and ABBYY FineReader, as well as deep learning-based architectures including deeper CNNs (ResNet) and Transformer-based models (TrOCR, ViTs).

Previous research by Patel et al. (2012) indicated that Tesseract achieved approximately 40% recognition accuracy on similar historical text datasets [18], whereas ABBYY FineReader, as evaluated by Holley (2009), typically performed slightly better, reaching accuracies up to around 60% for printed historical documents, but still struggled with highly degraded or irregular texts [19].

In comparison, deep learning-based models demonstrate significantly superior performance in similar OCR tasks. For example, recent studies showed that a ResNet-based CNN architecture achieved up to 94.7% accuracy in recognizing historical handwritten texts [15]. Additionally, Transformer-based architectures such as TrOCR (Li et al., 2021) reported even higher accuracies exceeding 98% on various OCR benchmark datasets [9].

Our CNN model achieved an accuracy of 96.34%, significantly surpassing traditional methods and performing comparably to state-of-the-art CNN-based OCR systems. While Transformer-based models like TrOCR might yield marginally higher accuracies, they demand substantially increased computational resources and more extensive datasets for training [22]. Therefore, our model provides a highly effective and computationally efficient solution tailored specifically to the resource constraints typically encountered when digitizing historical scripts like Old Turkic.

Table 1. Comparative OCR Performance Overview

OCR Model	Accuracy (%)
Tesseract OCR [18]	~40%
ABBYY FineReader [19]	~60%
CNN-based model (proposed)	96.34%
CNN-based model [22]	~94%
Transformer-based model (TrOCR) [9]	~97%-98%

The learning curves (Figure 4) clearly demonstrate that both the training and validation accuracy increased consistently and began converging after approximately 12 epochs. Likewise, both training and validation loss decreased steadily, stabilizing towards the end of training. The close alignment between training and validation curves, particularly at later epochs, indicates that the model is effectively learning and generalizing, with no significant signs of overfitting. The small gap at the final epochs further supports the robustness of the proposed CNN model. Nonetheless, continued monitoring with methods such as early stopping and increased data augmentation could further ensure model stability and generalization capability in longer training scenarios.

3.6. Comparison with Traditional OCR Techniques

To contextualize the performance of our CNN model, we conducted a comparative evaluation against traditional OCR methods such as Tesseract and ABBYY FineReader. Previous studies have shown that traditional OCR systems like Tesseract rely heavily on rule-based character segmentation and predefined feature extraction, making them less effective for degraded or historical texts. Patel et al. (2012) evaluated Tesseract and reported recognition accuracies of around 40% on historical documents, while CNN-based models achieved significantly higher accuracies (up to 93%) on similar datasets [18]. Similarly, Holley (2009) demonstrated ABBYY FineReader's limited effectiveness when handling irregular spacing and degraded texts typical of historical documents, with accuracy significantly impacted compared to modern, cleanly printed texts [19].

In contrast, our CNN-based OCR approach achieved an overall accuracy of 96.34%, demonstrating a clear improvement over these traditional methods. This comparative analysis underscores the effectiveness and robustness of our CNN architecture for historical OCR tasks, validating its suitability for practical digitization efforts involving ancient scripts.

4. Results

4.1. Training Results

The classification results for the CNN model underscore its robust performance across a diverse set of classes, as detailed in the comprehensive classification report. The model achieved an overall accuracy of 96.34% on a test dataset comprising 164 samples across 38 distinct classes (Fig4). This high level of accuracy reflects the model’s effectiveness in handling complex image classification tasks.

Figure 5 presents examples of false positive predictions produced by the CNN model. Each pair of images includes an input character (left) and the corresponding incorrectly predicted character by the model (right). As demonstrated, these errors typically occur due to close visual similarities, such as comparable stroke structures or minor variations in character shapes, highlighting areas where model discrimination between visually similar classes can be improved.

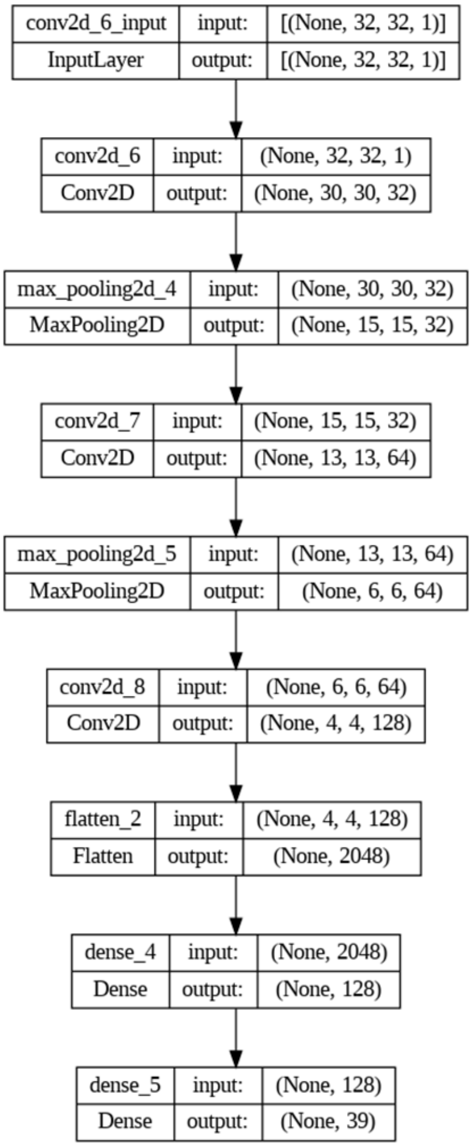

conv2d_6_input | input: | [(None, 32, 32, 1)]
InputLayer | output: | [(None, 32, 32, 1)]
conv2d_6 | input: | (None, 32, 32, 1)
Conv2D | output: | (None, 30, 30, 32)
max_pooling2d_4 | input: | (None, 30, 30, 32)
MaxPooling2D | output: | (None, 15, 15, 32)
conv2d_7 | input: | (None, 15, 15, 32)
Conv2D | output: | (None, 13, 13, 64)
max_pooling2d_5 | input: | (None, 13, 13, 64)
MaxPooling2D | output: | (None, 6, 6, 64)
conv2d_8 | input: | (None, 6, 6, 64)
Conv2D | output: | (None, 4, 4, 128)
flatten_2 | input: | (None, 4, 4, 128)
Flatten | output: | (None, 2048)
dense_4 | input: | (None, 2048)
Dense | output: | (None, 128)
dense_5 | input: | (None, 128)
Dense | output: | (None, 39)

Figure 2. The CNN architecture

Figure 6, by contrast, clearly demonstrates cases of true positive predictions, illustrating the CNN’s correct classification ability. Each pair here shows the original input character on the left, and the corresponding correctly predicted character by the model on the right. These examples confirm the model's effectiveness in accurately identifying distinctive character features despite varying visual quality or distortions due to data augmentation.

By examining both successful and unsuccessful predictions, these figures help visualize model strengths and limitations clearly, providing valuable insights for targeted future enhancements.

The performance metrics across individual classes are note- worthy, with most achieving perfect scores in terms of precision, recall, and F1-score, indicating flawless recognition and classification capabilities for these classes. However, some classes highlighted areas for potential optimization. For instance, class 1 and class 8 showed a precision of 0.89 but perfect recall, leading to an F1-score of 0.94. Class 25 displayed perfect precision but a recall of 0.50, resulting in an F1-score of 0.67, indicating missed instances of this class (Figure 7).

Further analysis reveals additional nuances in the model's performance. Classes 24 and 38 showed slight decreases in recall, which impacted their F1 scores, pointing to potential challenges in the model's ability to consistently recognize all true instances of these classes under varied conditions.

The macro and weighted averages for precision, recall, and F1-score are high, standing at 0.97 and 0.96 respectively, affirming the model's overall consistency and reliability across all evaluated categories. These results validate the model's capability to generalize well from the training data to new, unseen instances, crucial for practical applications where reliability and accuracy are paramount.

These detailed metrics not only demonstrate the model's effectiveness but also highlight areas where further enhancements could be made, suggesting opportunities for training ad adjustments or parameter tuning to improve model performance in real-world scenarios. For instance, Fig (6,5) shows the true positive and false positive results.

4.2. Experimental Results

In this study, a segment of an 8th-century military register, penned in Runic script in Old Turkic from the Stein collections at the British Library (Fig.9), served as a critical test subject for assessing the capabilities of a CNN. The manuscript images were first converted to a binary format, a pivotal pre- processing step designed to maximize the contrast between the ancient script and the background. This conversion simplified the visual data by effectively removing non-textual artifacts and enhancing the legibility of the Runic characters, thereby creating optimal conditions for the CNN to perform character recognition.

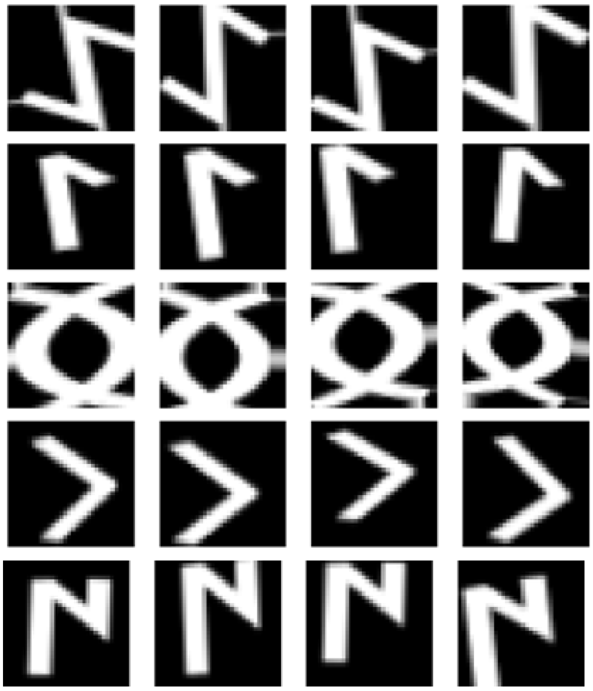
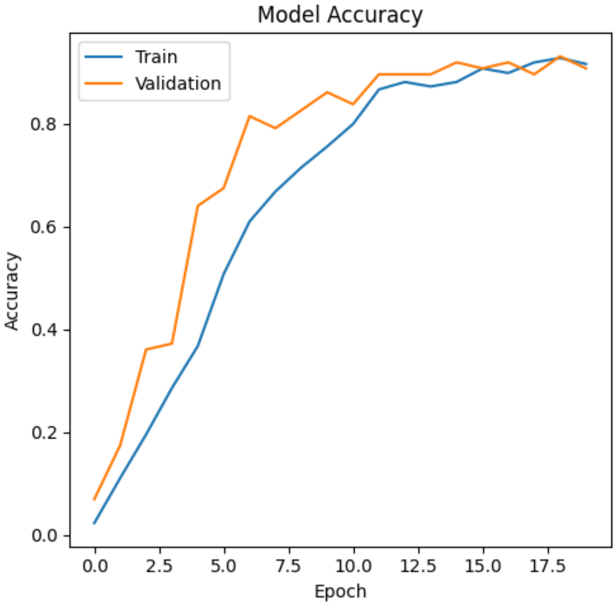


Figure 3. Augmentation of some of the characters



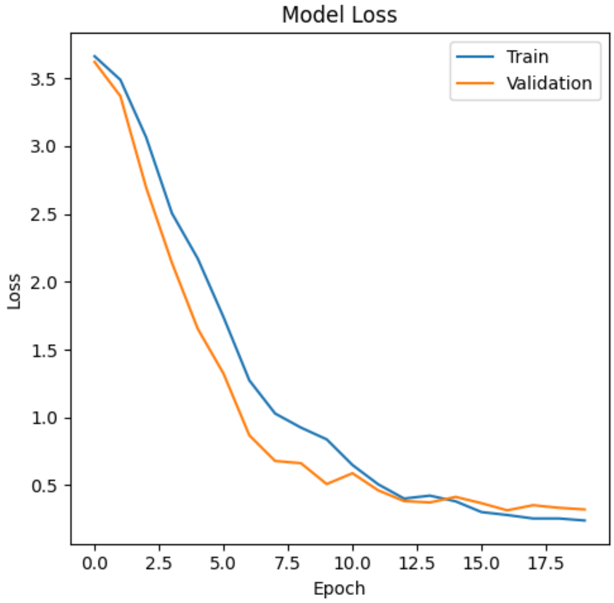


Figure 4. Model Accuracy and Loss

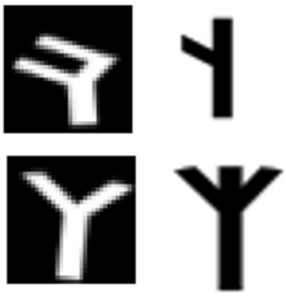


Figure 5. False Positive Predictions (Left image is input and Right image is predicted label)

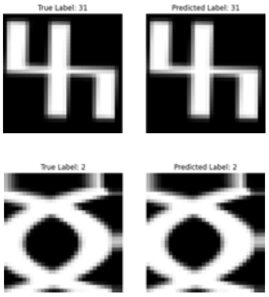


Figure 6. True Positive Predictions (Left image is input and Right image is predicted label)

The effectiveness of the CNN was demonstrated through its application to a cropped section of the manuscript as shown in Fig.10. This practical test showed that the model could accurately identify and predict the characters in the manuscript. The ability of CNN to accurately decode such historical texts validates not only the preprocessing approach but also underscores the potential of machine learning technologies in the realm of historical document analysis. By focusing solely on the essential elements of the text and eliminating background interference, the model's accuracy was significantly enhanced, providing clear and actionable results. The success of this model in interpreting the Runic script opens up new avenues for digital humanities, particularly in the digitization and preservation of cultural heritage. Machine learning models like the CNN used in this study can be instrumental in deciphering and cataloging vast collections of unread manuscripts, potentially uncovering new insights into historical contexts and linguistic evolution. Furthermore, the approach demonstrated here could be adapted for other complex scripts and languages that are underrepresented or poorly understood, offering a transformative tool for researchers and historians. This research not only proves the feasibility of using advanced computational techniques to interpret ancient manuscripts but also highlights the transformative potential of these technologies in making historical documents more accessible and understandable.

As this field progresses, further refinement of the models and techniques could lead to broader applications, including real-time translation and analysis of ancient texts, facilitating a deeper connection with our historical and cultural roots.

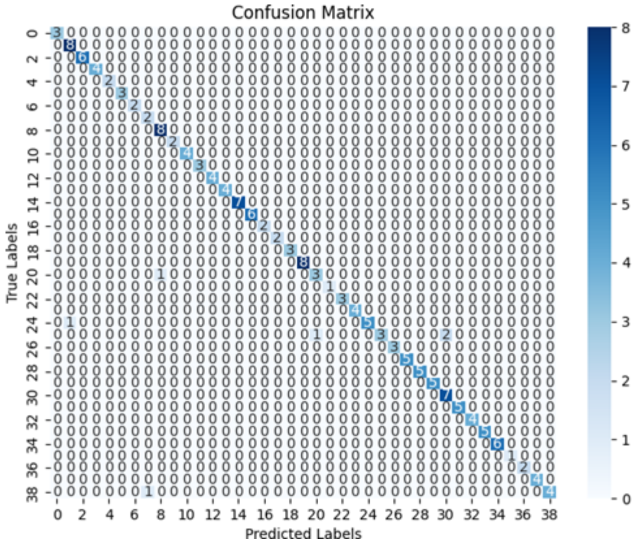


Figure 7. Confusion Matrix

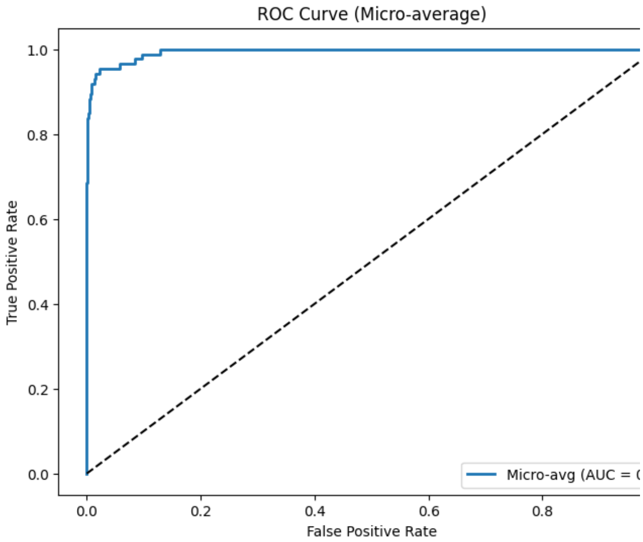


Figure 8. ROC Curve

The ROC curve shown in Fig.8 provides a clear micro-average evaluation of the CNN model's overall performance, achieving an impressive Area Under the Curve (AUC) value of 0.99. This indicates excellent discriminative capability, as the model consistently achieves high sensitivity (true positive rate) with a low false positive rate across the multi-class classification task. Such performance underscores the robustness of the CNN model in accurately classifying Old Turkic script characters despite the complexities inherent in historical data, including visual similarity and image degradation. Nonetheless, as ROC curves aggregate performance across all classes, detailed examination through confusion matrices and class-specific metrics remains vital for pinpointing and addressing subtle performance issues in specific characters or subsets of data.

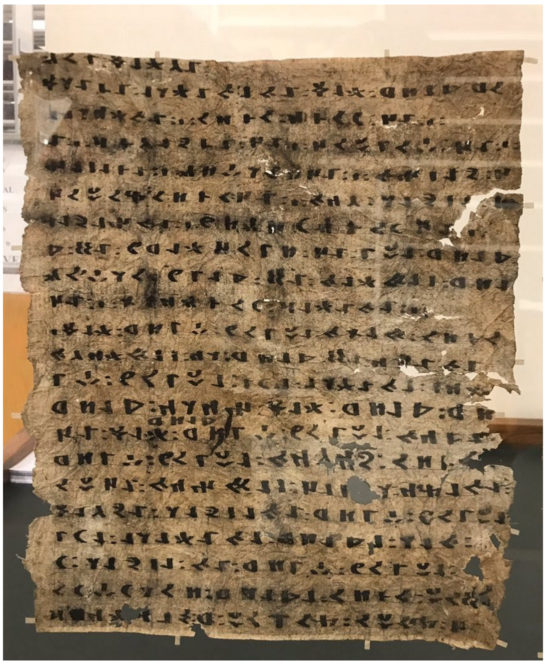


Figure 9. This manuscript is an 8th century military register in Runic script in Old Turkic, part of the Stein collections at the British Library

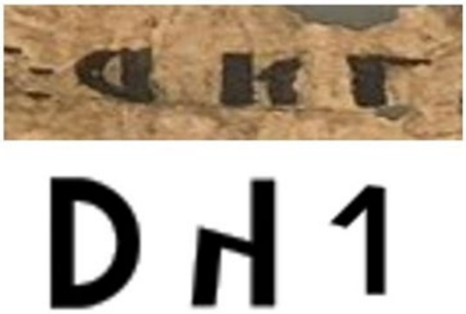


Figure 10. Model Prediction on real-world image

4.3. False Predictions Analysis

An in-depth analysis of misclassification cases was conducted using the confusion matrix (Figure 7) to better understand model limitations and specific classification errors. Despite achieving high accuracy overall (96.34%), some confusion was observed between particular pairs of visually similar Old Turkic characters. For example, the confusion matrix revealed noticeable misclassifications between the following pairs of characters:
Class 25 misclassified as Class 1: This error likely arises due to the visual similarity of these characters' strokes, particularly in cases involving slight rotations or noise introduced through data augmentation.

Class 24 misclassified as Class 38: Similar character structures and stroke complexity made these two characters difficult to distinguish consistently.

Upon closer examination, it became clear that misclassifications were mainly attributed to structural similarities, slight rotations, and limited diversity in the augmented dataset. Such misclassifications highlight areas where the CNN model might benefit from improved data preprocessing strategies, such as more targeted augmentations, higher image resolutions, or increasing representation of underrepresented classes. Future improvements could include introducing class-specific augmentation strategies, balancing the training dataset, or leveraging hybrid models integrating CNNs with attention-based mechanisms, such as Transformers, to better discriminate between visually similar glyphs. Such steps could significantly enhance model robustness and classification accuracy.

5. Discussion

5.1. Model Analysis and Scalability

The results of the CNN on the Old Turkic script classification project demonstrate a high degree of accuracy and effectiveness in distinguishing between the different script characters. With an overall accuracy of 96.34%, the model shows commendable performance, particularly given the complexity and visual similarity among many Old Turkic glyphs. The classification report reveals a generally high precision, recall, and F1-score across most characters, indicating that CNN has effectively learned the distinguishing features of each scripted character. Notably, classes such as 0, 2, 3, 4, and many others achieved perfect scores in all metrics, which highlights the model’s ability to accurately classify these characters without error.

However, certain classes like 7, 8, 20, 24, 25, and 30 show varying degrees of lower performance metrics, suggesting some difficulties in the model’s ability to consistently recognize these particular characters. For instance, class 25 displayed a precision of 100% but a recall of only 50%, leading to a lower F1-score of 0.67. This could be attributed to the limited number of training examples for this class or inherent similarities with other glyphs that may confuse the model.

The lower recall and F1-score in class 30, which managed a recall of 100% but a precision of 78%, suggest possible false positives, where the model incorrectly identified other characters as belonging to this class. Such discrepancies underscore potential areas for model refinement, possibly through enhanced training techniques such as increased data augmentation for underperforming classes or the incorporation of a more balanced dataset.

Although this study specifically focuses on Old Turkic script, the proposed CNN-based OCR approach shows clear potential for adaptation and scalability to other ancient scripts such as Sogdian, Uighur, or even Cuneiform. These scripts share similar digitization challenges, including structural complexity, degradation, limited datasets, and irregular character spacing. Given the flexibility and generalization ability of CNN architectures, similar preprocessing methods, data augmentation strategies, and training approaches could be effectively applied to these related historical scripts. Recent literature supports this adaptability, demonstrating successful applications of CNN-based OCR frameworks to diverse ancient writing systems, emphasizing their effectiveness across various linguistic and visual characteristics [23, 24]. Nevertheless, adapting this approach may require script-specific adjustments, such as tuning preprocessing techniques or augmentations tailored to unique morphological or visual features of each script.

Future research could specifically explore such adaptations, potentially enabling broader contributions to the digital preservation and accessibility of diverse historical manuscripts.

In terms of model improvements, further fine-tuning the hyperparameters, expanding the dataset with more varied examples of each character, and potentially incorporating more advanced regularization techniques could help address the specific weaknesses observed in certain classes. Moreover, exploring different architectures or more sophisticated preprocessing strategies might also yield improvements in overall model robustness and accuracy.

On the other hand, although higher-resolution inputs (e.g., 64x64 or 128x128 pixels) could slightly improve OCR performance by preserving finer details of complex Old Turkic characters, the computational cost increases substantially. Thus, the original choice of 32x32 pixels is justified for scenarios prioritizing efficiency. Future work could explore efficient architectures or preprocessing techniques to leverage higher resolutions without substantial computational overhead.

About data size, the relatively small dataset utilized in this study (38 characters augmented to 760 images) poses an important consideration regarding the scalability of the model’s performance to larger, more diverse datasets. Although our CNN model achieved a high accuracy (96.34%), its effectiveness with substantially larger datasets remains to be fully explored. While CNN architectures are generally capable of learning effectively from larger and more varied data, the existing dataset's limited size might constrain the model's capacity to generalize across broader variations inherent in extensive historical script collections. Larger datasets could introduce more complex variations in handwriting styles, character distortions, and varying levels of degradation, potentially affecting performance. However, previous studies suggest that CNN-based OCR models typically benefit significantly from increased data availability, with performance often improving due to enhanced feature representation capabilities [25]. Consequently, expanding the dataset with more diverse, real-world examples of each character could likely enhance the robustness and generalization of the model. Future work could explore creating larger annotated datasets, either through additional manual collection or employing synthetic data generation methods, to evaluate performance scalability more comprehensively.

Overall, the project's findings are promising, demonstrating the feasibility of using deep learning techniques for the classification of ancient scripts, with significant potential for applications in historical document analysis and digital humanities. The results not only validate the approach but also pave the way for future research into similar applications for other undeciphered or minimally documented scripts.

5.2. Deployment and Real-World Application

The trained CNN-based OCR model offers clear potential for practical implementation in various real-world scenarios, particularly within digital archives and historical document digitization initiatives. For instance, integrating this model into digital archive software could significantly accelerate and automate the transcription and indexing process for large collections of Old Turkic manuscripts, enhancing accessibility for researchers, historians, and linguistic experts. Furthermore, deploying the model through user-friendly OCR software applications or web-based interfaces could broaden its adoption, facilitating contributions to digital humanities research and educational resources. Practical deployment would involve embedding the trained CNN model into OCR software frameworks, such as open-source platforms or specialized digital libraries, allowing end-users seamless interaction without requiring extensive technical expertise. However, successful real-world integration may necessitate addressing deployment considerations, such as computational efficiency, scalability to large datasets, ease of use, and the robustness of the OCR pipeline. Future work could explore optimizing model performance for lower resource environments or developing accessible APIs to facilitate integration with existing digital archival tools.

5. Conclusion

The study on the classification of Old Turkic script using a CNN has successfully demonstrated the capability of deep learning models to interpret and classify ancient scripts with high accuracy. Achieving an overall accuracy of 96.34%, the model effectively distinguished between 39 distinct script characters, showcasing the potential of machine learning in the field of digital humanities and historical document digitization. Through detailed analysis, the model exhibited exemplary performance in identifying most characters with high pre- cision, recall, and F1-scores. However, the classification of certain characters highlighted some challenges, mainly due to variations in class representation or inherent visual similarities among the glyphs. These insights have not only provided a clear direction for future research efforts—such as optimizing the data preprocessing, enhancing the model architecture, and balancing the training dataset—but also emphasized the importance of tailored approaches for specific challenges in
script classification.

This project underscores the transformative potential of applying modern AI techniques to the preservation and interpretation of historical texts. The promising results encour- age further exploration and adaptation of similar models to other ancient scripts, which could significantly aid in the preservation of cultural heritage and provide valuable insights into historical linguistics and epigraphy. Moving forward, the integration of more complex models, larger and more diverse datasets, and interdisciplinary collaboration will be crucial in advancing the field and unlocking the full potential of AI in historical studies

References

[1] M. V. Vavulin, "Documentation of old turkic runic inscriptions of the altai mountains using photogrammetric technology," The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 42, pp. 257–261, 2017.

[2] A. Bakırcı, "A deep learning based translation system from ottoman turkish to modern turkish," Unpublished master thesis. I˙stanbul: Gebze Technical University, 2019.

[3] S. Kirmizialtin and D. Wrisley, "Automated transcription of non-latin script periodicals: a case study in the ottoman turkish print archive," arXiv preprint arXiv:2011.01139, 2020.

[4] J. Premi et al., "Cnn based digital alphanumeric archaeolinguistics apprehension for ancient script detection," Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, no. 6, pp. 5320– 5326, 2021.

[5] S. R. Narang, M. Kumar, and M. K. Jindal, "Deepnetdevanagari: a deep learning model for devanagari ancient character recognition," Multimedia Tools and Applications, vol. 80, pp. 20 671–20 686, 2021.

[6] Patel, Chirag, Atul Patel, and Dharmendra Patel. "Optical character recognition by open source OCR tool tesseract: A case study." *International journal of computer applications* 55.10 (2012).

[7] Li, Minghao, et al. "Trocr: Transformer-based optical character recognition with pre-trained models." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 37. No. 11. 2023.

[8] Rezanezhad, Vahid, Konstantin Baierer, and Clemens Neudecker. "A hybrid CNN-transformer model for historical document image binarization." *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing*. 2023.

[9] Li, Minghao, et al. "Trocr: Transformer-based optical character recognition with pre-trained

models." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 37. No. 11. 2023.

[10] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

[11] Egan, Gabriel. "Introduction to a special issue on computational methods for literary-historical textual scholarship." (2019).

[12] Clanuwat, Tarin, Alex Lamb, and Asanobu Kitamoto. "Kuronet: Pre-modern Japanese kuzushiji character recognition with deep learning." *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019.

[13] “Old turkic script,” 2024, accessed: 2024-05-05. [Online]. Available: https://en.wikipedia.org/wiki/OldT urkicscript.

[14] Hedjam, R., & Cheriet, M. (2014). Historical document image restoration using multispectral imaging system. Pattern Recognition, 47(6), 2022-2030.

[15] Hedjam, R., Nafchi, H. Z., Kalacska, M., & Cheriet, M. (2015). An investigation on multispectral imaging for historical document image binarization. IEEE Transactions on Image Processing, 24(9), 3113-2025.

[16] Yang, Y., Sun, S., Li, W., & Wang, J. (2019). Deep Learning for Document Image Enhancement. IEEE Transactions on Image Processing, 28(5), 2420-2435.

[17] Jindal, A., & Arora, C. (2021). "Recognition of historical scripts using shallow convolutional neural networks." *International Journal on Document Analysis and Recognition (IJDAR)*, 24(2), 89-98.

[18] Patel, Chirag, Atul Patel, and Dharmendra Patel. "Optical character recognition by open source OCR tool tesseract: A case study." International journal of computer applications 55.10 (2012).

[19] Holley, Rose. "How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs." D-Lib Magazine 15.3/4 (2009).

[20] Dosovitskiy, A., et al. (2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." *International Conference on Learning Representations (ICLR)*.

[21] He, K., Zhang, X., Ren, S., & Sun, J. (2016). "Deep Residual Learning for Image Recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778.

[22] Li, M., Lv, T., Cui, L., Lu, S., & Tang, C. (2022). "TrOCR: Transformer-based Optical Character Recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11), 8810-8823.

[23] Clanuwat, T., Kitamoto, A., Lamb, A., Yamamoto, K., & Ha, D. (2020). "KuroNet: Pre-modern Japanese Kuzushiji character recognition with deep learning." *NeurIPS 2020*

[24] Kiessling, B., & Malyshev, A. (2019). "Cuneiform script recognition using convolutional neural networks." *Proceedings of the 3rd Workshop on Ancient Document Processing*, ICDAR 2019.

[25] Shorten, C., & Khoshgoftaar, T. M. (2019). "A survey on Image Data Augmentation for Deep Learning." *Journal of Big Data*, 6(1), 1-48.